"Using Data Mining Techniques For Customer Engagement"

Minor Project Report Submitted

To

Chhattisgarh Swami Vivekananda

Technical University, Bhilai (C.G.), India

for

The award of the degree

of

BACHELOR OF TECHNOLOGY(Hons.)

In

COMPUTER SCIENCE & ENGINEERING
(Artificial Intelligence / Data Science)

By

Geetanshu Dev Meshram
B.Tech. 5th Semester
Roll No. 300012821043
Enrollment No. CB4691

Under the Guidance of

K.Vibhooti Rajkumar

Assistant Professor
Computer Science and Engineering
UTD,CSVTU BHILAI (C.G.)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UTD, CSVTU BHILAI (C.G.)
Session 2023-2024

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UTD, CSVTU,BHILAI(C.G.)

DECLARATION BY THE CANDIDATE

I, the undersigned solemnly declare that the thesis entitled **"Using Data Mining Techniques For Customer Engagement"** is based on my work carried out during the course of my study under the supervision of **K.Vibhooti Rajkumar,** Assistant professor of the Department of Computer Science and Engineering, Chhattisgarh Swami Vivekanand Technical University,UTD,Bhilai(C.G.), India.

_____

Geetanshu Dev Meshram

Roll No. 300012821043

Enrollment No. CB4691

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UTD, CSVTU,BHILAI(C.G.)

CERTIFICATE OF THE SUPERVISOR

This is to certify that the incorporation in the thesis "**Using Data Mining Techniques For Customer Engagement"** is a record of research work carried out by Geetanshu Dev Meshram, bearing Roll No. 300012821043 **,**Enrollment No. CB4691 guidance and supervision for the award of the degree of **Bachelor of Technology in Computer Science & Engineering (Data Science)** of Chhattisgarh Swami Vivekanand Technical University,UTD, Bhilai (C.G.) India.

To the best of my knowledge and belief the thesis

I. Embodies the work of the candidate himself,

II. Has duly been completed in the specified time,

III. Fulfill the requirement of the Ordinance relating to the B.Tech.. degree of the University and

IV. Is up to the desired standard both in respect of contents and language for being referred to the examiners.

(Signature of H.O.D.)	(Signature of Supervisor)
Dr.Toran Verma	K.VibhootiRajkumar
Professor & HOD	AssistantProfessor
Department of CSE (AI/DS)	Department of CSE(AI/DS)

Forwarded to Chhattisgarh Swami Vivekananda Technical University, Bhilai (C.G.)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UTD, CSVTU,BHILAI(C.G.)

CERTIFICATE BY THE EXAMINERS

This is to certify that the project thesis entitled "**Using Data Mining Techniques For Customer Engagement"** was submitted by Geetanshu Dev Meshram student of B. Tech.(Hons) (CSE) (Roll No. 300012821043, Enrollment No. CB4691) has been examined by the undersigned as a part of the examination and is hereby recommended for the award of the degree of **Bachelor of Technology in Computer Science and Engineering (DataScience)** of Chhattisgarh Swami Vivekananda Technical University, UTD,Bhilai (C.G.), India.

_____ _____

Internal Examiner External Examiner

Date: Date:

ACKNOWLEDGEMENT

The real spirit of achieving a goal is through the way of excellence and serious discipline. I want to thank UTD,CSVTU, Bhilai for providing me with the necessary software, tools, and other resources to deliver my minor project work.

With gratitude and humanity, I acknowledge my indebtedness to **K.Vibhooti Rajkumar**, Assistant Professor UTD, CSVTU, under whose guidance I had the privilege to complete this project work. Also, I am grateful to all the faculty members of the department of CSE, who were always there at the need of the hour and provided me with all the help and facility, I required for the completion of my project work.

My greatest thanks go to my parents and family, who have been my driving force. My work would not be possible without their constant inspiration, encouragement, support, and love. Above all, I render my gratitude to the almighty, who bestowed self-confidence, Ability, and strength on me to complete this work.

Geetanshu Dev Meshram
Roll No. 300012821043
Enrollment No. CB4691

ABSTRACT

The evolution of Customer Engagement concept had a positive impact on how business and customer interact. The term Digital Customer Engagement emerged to empower such interaction via encouraging customers to use digital channels. Data mining techniques can help in identifying patterns, generating insights and making predictions for a massive amount of data. The purpose of this paper is to explore the current literature on data mining techniques to enhance digital customer engagement and review the impact of data mining on analyzing customer's attributes and its key performance indicators influenced by digital customer engagement and affecting business success.

The scope of the review encompasses the state of the art of scientific methodologies and models applied to identify customers with the highest potential towards digital engagement. A critical analysis also identified gaps in the literature. This study is the first to explicitly consider data mining techniques for enhancing digital customer engagement with a comprehensive analysis of customer's attributes in different domains. Keywords: Data Mining, Machine Learning, Literature, Digital Customer Engagement, Analysis and Algorithm.

Table of contents

List of Figures

List of Table

List of Abbreviations

1.	**KNN**	-	K Nearest Neighbour
2.	**CRM**	-	Customer Relationship management
3.	**PCA**	-	Principal component Analysis
4.	**ML**	-	Machine Learning
5.	**SVM**	-	Support Vector Machine
6.	**LDA**	-	Linear Discriminant Analysis

CHAPTER – I

INTRODUCTION

CHAPTER – I:

INTRODUCTION

1.1 Overview

In the rapidly evolving realm of business, the significance of customer engagement stands as a linchpin in determining the success and sustainability of enterprises. Beyond the traditional transactional nature of commerce, customer engagement represents a dynamic and reciprocal interaction between businesses and their clientele. It encompasses the multifaceted ways in which companies cultivate and maintain relationships with their customers, extending far beyond the point of purchase. As markets become more saturated and consumer expectations continue to soar, the ability of organizations to foster meaningful, ongoing connections with their customer base becomes not just advantageous but imperative.

Data mining techniques are powerful tools for analyzing large and complex datasets, and extracting useful insights and patterns from them. Data mining can be applied to various domains, such as marketing, health care, education, finance, and more. One of the most important applications of data mining is customer relationship management (CRM), which aims to enhance customer satisfaction, loyalty, retention, and profitability.

Customer engagement is a key aspect of CRM, as it reflects the degree of interaction and involvement between customers and a business. Customer engagement can be influenced by various factors, such as customer preferences, behavior, feedback, satisfaction, loyalty, and value. Data mining can help businesses understand and improve customer engagement by discovering hidden patterns, trends, and associations from customer data, and providing actionable insights and recommendations.

In this thesis, we will explore how data mining techniques can be used to support customer engagement in different scenarios and contexts. We will review the existing literature on data mining for customer engagement, and identify the main challenges and opportunities in this field. We will also propose and implement novel data mining methods and models for customer engagement, and evaluate their performance and effectiveness using real-world datasets. We will demonstrate how data mining can help businesses achieve their goals and objectives, and create value for both customers and themselves.

1.2 Thesis Goals and Objectives

The objective is to retain the loyalty for the company by improving customer engagement using Data Mining techniques.

1.3 Organization of Thesis

The rest of the thesis has been organized into four chapters. Following is a brief description of each chapter:

Chapter 2. Review of Related Work

This chapter deals with the different types of machine learning techniques used for customer engagement using different algorithms. This chapter also deals with a brief analysis of the techniques used for the better understanding of customer engagement using different methodologies done previously.

Chapter 3. Problem Identification

This chapter deals with the identification of the problem due to which we reached the solution and thought that this project helped in resolving the problem to an extinct.

Chapter 4. Proposed Methodology

This chapter deals with the methodology and techniques used in building the project with a proper workflow diagram.

Chapter 5. Result & Discussion

Here we mentioned the result and gave a brief discussion on we are solving the problem with result accuracies.

Chapter 6.Conclusion&Futurescope

This chapter deals with the conclusion of whether the problem is actually resolved or not and how much we can improve it further and also adds the future scope of what we can add to enhance its performance.

CHAPTER – II

LITERATURE REVIEW

CHAPTER II:

LITERATURE REVIEW

2.1 Paper Reviewed

The authors have attempted to classify papers in terms of the application of DM in various industrial domains.

While the authors carried out an in-depth analysis of DM technology, Table 1 lists only 35 relevant papers, all of which heavily emphasises the application of DM techniques in various industries. The remaining papers are algorithmic in nature and purely focus on CRM applications without specifically mentioning CRM; however, they too would come under the general category of DM and CRM. This category includes papers that discuss the benefits and difficulties of incorporating DM into CRM. The authors found that several papers suggest models that can be used in popular business domains and also contain advice on how this information can be extracted and utilised. The papers using tools for the analyses of data and proposes specific systems to business analysts are categorised under CRM tools.

Feng et al. (2008) exercised into the benefits of the CRM and DM combination. It is largely beneficial, since understanding the customer ensures their proper needs are met. Ngai et al. (2009) reviewed DM techniques and its application in CRM, principles, customer acquisition, customer retention and application of DM techniques. Dempster et al. (2008) statistical research paper discussed statistical techniques and calculus in research results. Zvireliene et al. (2009) designed and explained a detailed CRM system based on DM. Various DM techniques can supplement numerous CRM functions. Pan et al. (2007) created a pre-processing framework which helps in predicting customer's switching to competitors. A framework has been applied in software system and benchmarked. It accurately predicted customer's likelihood of leaving and thus gives the company a change to act on them. Ranjan and Malik (2007) proposed an educational model based on DM techniques. The model viewed the effect of the processes related to admission, course delivery and recruitments. The paper educational processed and DM effect on the educational model. Adela et al. (2011) paper explored the application of optimisation techniques (OT) on DM and certain predictive model's efficacy. Models are useful in helping managers take better decisions in time. The customers are appropriately used to improve business. OT with DM should be researched further.

2.2 Summary

In the present chapter, the literature survey made so far related to the work has been briefly discussed.

S.no	Year	Paper Title	Journals	Research Finding
1.	2020	Predicting customer's gender and age depending on mobile phone data.	Authored byWassouf et al. The article is published in the Journal of Big Data, volume 7, issue 29 in 2020.	Data mining improves customer engagement by personalizing, retaining, and targeting customers.Data mining faces limitations due to data quality, privacy, and currency issues.
2.	2021	Customer Relationship Management Using Data Mining Model.	Authored by S.Balamuruga n and Dr. M. Selvalakshmi	Data mining improves customer engagement by personalizing, retaining, and targeting customers. Data mining faces limitations due to data quality, currency, and complexity.

3.	2008, 2010, 2011, 2012.	DM and CRM in general .	Jain et al. (2008), Zhao (2011), Pabreja and Datta (2012), Kumar et al. (2012), Zhang (2010), Ngai et al. (2009), Liao et al. (2012), Burt et al. (2011) and Verhoef et al. (2010)	The other categories represent industries suggesting models or actual algorithms to use DM to maximise business processes.

CHAPTER – III

PROBLEM IDENTIFICATION

Chapter III:

PROBLEM IDENTIFICATION

- The application of data mining techniques in customer engagement, while holding significant potential for personalized interactions and improved business outcomes, is confronted with several critical challenges that impede its seamless integration and optimal utilization. The following issues have been identified as key obstacles in realizing the full potential of data mining for customer engagement:
- The reliability and accuracy of customer data are fundamental to the success of data mining initiatives. Inconsistent, incomplete, or inaccurate data can compromise the effectiveness of mining techniques and result in flawed insights, hindering the ability to engage customers meaningfully.
- Integrating data mining techniques seamlessly with existing customer relationship management (CRM) systems and organizational workflows poses a substantial challenge. Many businesses struggle with the compatibility of data mining tools with legacy systems, leading to inefficiencies in implementation and utilization.
- The adoption of data-driven decision-making in customer engagement often encounters resistance within organizational cultures. Employees may be hesitant to embrace new methodologies, and a lack of data literacy can hinder the successful integration of data mining techniques into daily operation.
- Successful implementation of data mining requires substantial computational resources, and scalability becomes a critical concern as datasets grow. Many organizations may face challenges in allocating the necessary resources and infrastructure to support the evolving demands of data mining for customer engagement.
- The dynamic nature of customer behavior requires real-time responsiveness in engagement strategies. Data mining techniques must be capable of providing timely insights to enable organizations to adapt quickly to changing customer preferences, posing a challenge in terms of algorithmic efficiency and computational speed.
- Addressing these identified problems is imperative to unlock the full potential of data mining techniques in customer engagement and to establish a foundation for organizations to build meaningful, long-lasting relationships with their customer.

CHAPTER – IV

PROPOSED METHODOLOGY

Chapter IV:
METHODOLOGY

Dataset description:

This dataset is having the data of 1K+ Amazon Product's Ratings and Reviews as per their details listed on the official website of Amazon.

- discount_percentage - Percentage of Discount for the Product
- rating - Rating of the Product
- rating_count - Number of people who voted for the Amazon rating
- about_product - Description about the Product
- user_id - ID of the user who wrote review for the Product
- user_name - Name of the user who wrote review for the Product
- review_id - ID of the user review
- review_title - Short review
- review_content - Long review
- img_link - Image Link of the Product
- product_id - Product ID
- product_name - Name of the Product
- category - Category of the Product
- discounted_price - Discounted Price of the Product
- actual_price - Actual Price of t
- Amazon is an American Tech Multi-National Company whose business interests include E-commerce, where they buy and store the inventory, and take care of everything from shipping and pricing to customer service and returns. I've created this dataset so that people can play with this dataset and do a lot of things as mentioned below
- Dataset Walkthrough

- Understanding Dataset Hierarchy
- Data Preprocessing
- Exploratory Data Analysis
- Making Recommendation System

	product_id	product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about_product	
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	₹399	₹1,099	64%	4.2	24,269	High Compatibility : Compatible With iPhone 12...	AG3D6O4STAQKAY2UVGEUV46K
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹349	43%	4.0	43,994	Compatible with all Type C enabled devices, be...	AECPFYFQVRUWC3KGNLJIOR
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹1,899	90%	3.9	7,928	【Fast Charger& Data Sync】-With built-in safet...	AGU3BBQ2V2DDAMOAKGFAWD
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	₹329	₹699	53%	4.2	94,363	The boAt Deuce USB 300 2 in 1 cable is compati...	AEWAZDZZJLQUYVOVGBEUKS
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	₹154	₹399	61%	4.2	16,905	[CHARGE & SYNC FUNCTION]- This cable comes wit...	AE3Q6KSUK5P75D5HFYHCRA

Fig.4.1:visualization of Dataset

Data preprocessing:

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

Data Reduction: This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

Data Discretization: This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

Data Normalization: This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

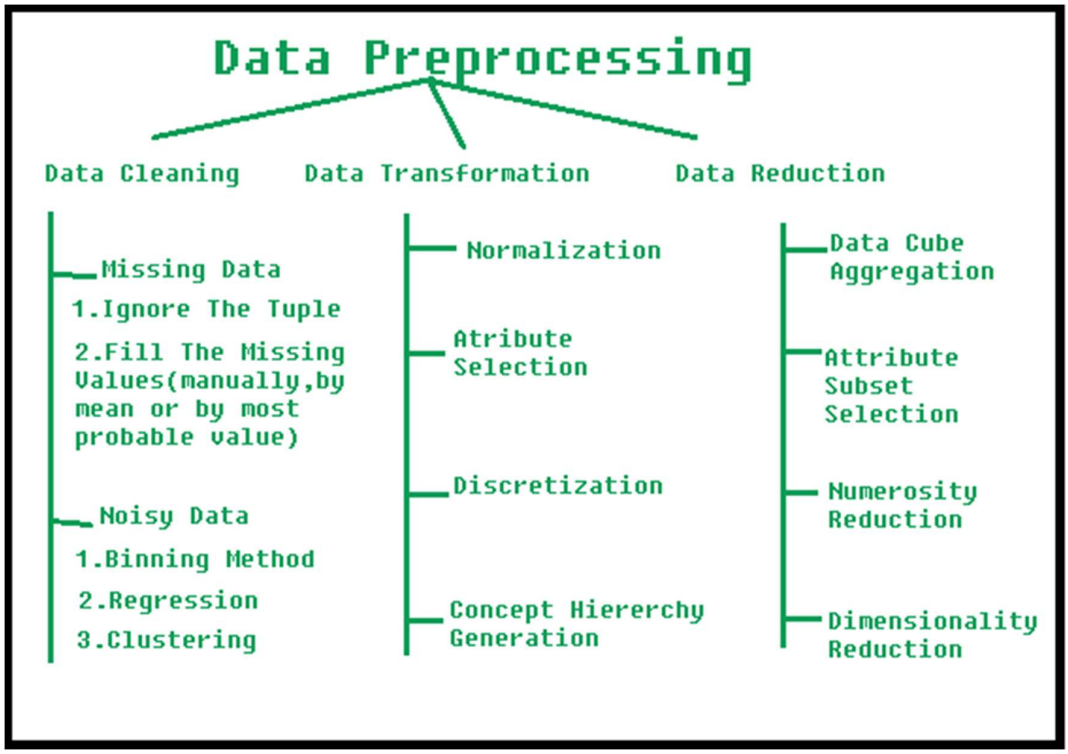

Fig.4.2:flowchart of datapreprocessing

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**

  This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1.Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2.Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1.Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

2.Regression:
Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

3.Clustering:
This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters

Data Transformation:
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

- **Normalization:**
  It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

- **Attribute Selection:**
  In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

- **Discretization:**
  This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

- **Concept Hierarchy Generation:**
  Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:
Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

Sampling: This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

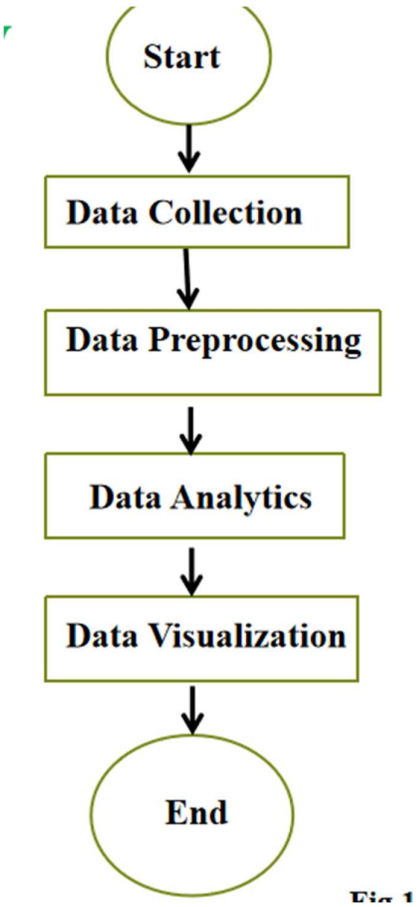

Fig.4.3:workflow diagram

Machine Learning techniques used:

1. Naive Bayes:

Naive Bayes is a family of classification algorithms based on Bayes' theorem, which is a mathematical formula that relates the conditional and marginal probabilities of two events. Naive Bayes classifiers assume that the features of the data are independent of each other, given the class label. This means that the effect of a feature on the prediction does not depend on the values of other features. This is a simplifying assumption that makes the computation easier, but it may not always hold true in real-world data.

Naive Bayes classifiers work by calculating the posterior probability of each class, given the features of the data, and then choosing the class with the highest probability. The posterior probability is computed using Bayes' theorem, which can be written as:

$P(C|F)=P(F)P(F|C)P(C)$

Where:

- P(C|F) is the posterior probability of class C, given the features F.
- P(F|C) is the likelihood probability of features F, given the class C.
- P(C) is the prior probability of class C.
- P(F) is the marginal probability of features F.

To calculate the posterior probability of each class, we need to estimate the likelihood, prior, and marginal probabilities from the data. Depending on the type and distribution of the features, different methods can be used to estimate these probabilities. For example, if the features are continuous and normally distributed, we can use the mean and standard deviation of each feature for each class to calculate the likelihood probability. If the features are discrete and multinomial, we can use the frequency counts of each feature value for each class to calculate the likelihood probability.

Naive Bayes classifiers are simple, fast, and effective for many classification tasks, especially for text and document classification. They can handle high-dimensional data and deal with missing values by ignoring them. However, they may not perform well if the independence assumption is violated or if some features are correlated with each other. They may also suffer from zero-frequency problem, which occurs when a feature value has never been observed with a class in the training data, resulting in a zero likelihood probability.

2. Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

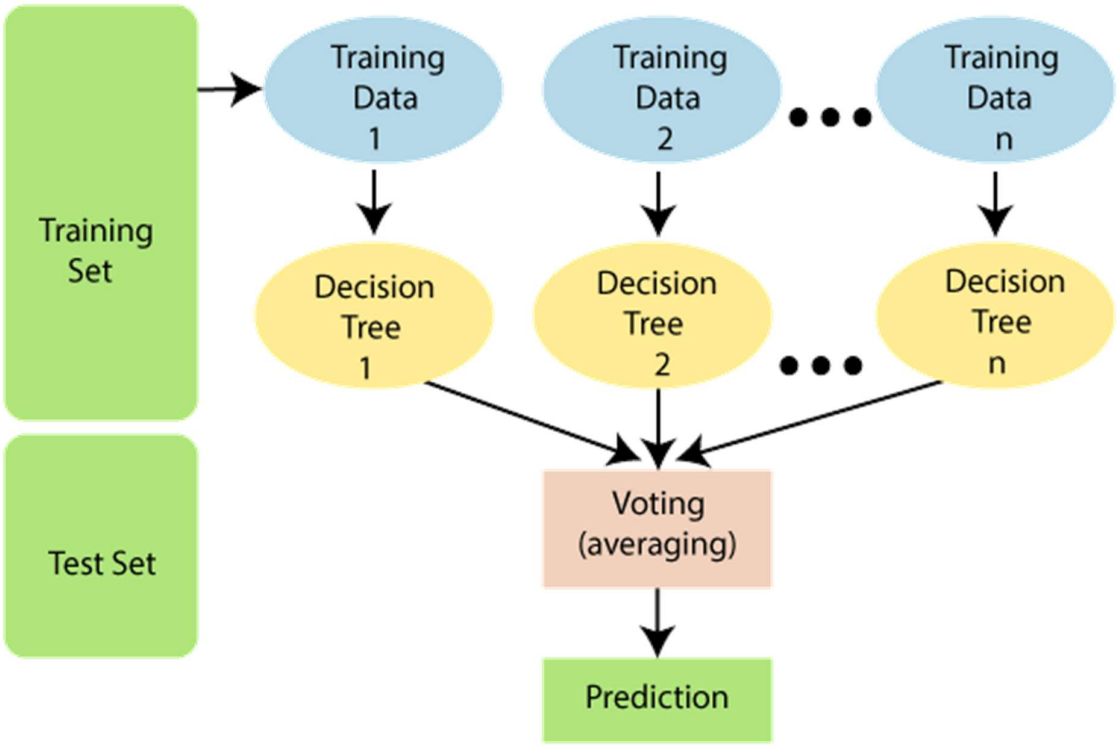

Fig.4.4:explanation of Random forest

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier.

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.K-Nearest Neighbour:

The K-Nearest Neighbors (KNN) algorithm is a robust and intuitive machine learning method employed to tackle classification and regression problems. By capitalizing on the concept of similarity, KNN predicts the label or value of a new data point by considering its K closest neighbours in the training dataset. In this article, we will learn about a supervised learning algorithm (KNN) or the k – Nearest Neighbours, highlighting it's user-friendly nature.

What is the K-Nearest Neighbors Algorithm?

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features:

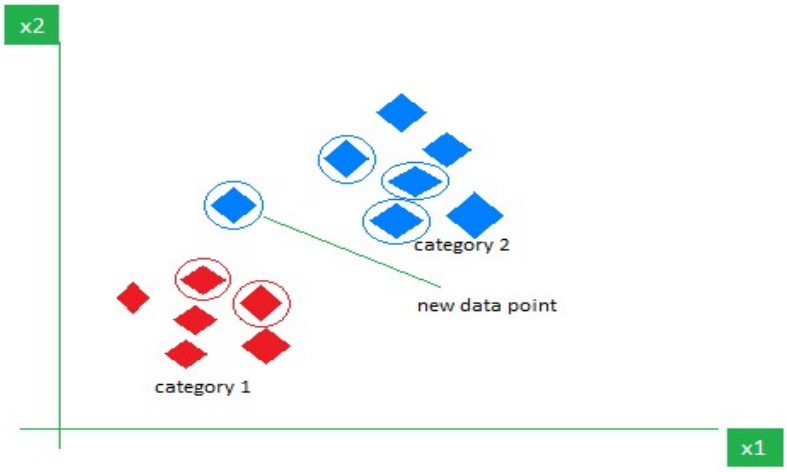

fig.4.5:clustering formation

Now, given another set of data points (also called testing data), allocate these points to a group by analyzing the training set. Note that the unclassified points are marked as 'White'.

Intuition Behind KNN Algorithm

If we plot these points on a graph, we may be able to locate some clusters or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbors belong to. This means a point close to a cluster of points classified as 'Red' has a higher probability of getting classified as 'Red'.

Intuitively, we can see that the first point (2.5, 7) should be classified as 'Green' and the second point (5.5, 4.5) should be classified as 'Red'.

Why do we need a KNN algorithm?

(K-NN) algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution. It can also handle both numerical and categorical data, making it a flexible choice for various types of datasets in classification and regression tasks. It is a non-parametric method that makes predictions based on the similarity of data points ina given dataset. K-NN is less sensitive to outliers compared to other algorithms.

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data.

Distance Metrics Used in KNN Algorithm

As we know that the KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. For this purpose, we use below distance metrics:

Euclidean Distance

This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2}]$$

Manhattan Distance

Manhattan Distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions.

Minkowski Distance

We can say that the Euclidean, as well as the Manhattan distance, are special cases of the Minkowski distance.

$$d(x, y) = (\sum_{i=1}^{n} (x_i - y_i)^p)^{\frac{1}{p}}$$

From the formula above we can say that when p = 2 then it is the same as the formula for the Euclidean distance and when p = 1 then we obtain the formula for the Manhattan distance.

The above-discussed metrics are most common while dealing with a Machine Learning problem but there are other distance metrics as well like Hamming Distance which come in handy while dealing with problems that require overlapping comparisons between two vectors whose contents can be boolean as well as string values.

4.Decision Tree:

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

What is a Decision Tree?

A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very

powerful algorithms. And it is also used in Random Forest to train on different subsets of training data, which makes random forest one of the most powerful algorithms in machine learning.

Decision Tree Terminologies

Some of the common Terminologies used in Decision Trees are as follows:

- **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
- **Decision/Internal Node**: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
- **Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Parent Node:** The node that divides into one or more child nodes.
- **Child Node:** The nodes that emerge when a parent node is split.
- **Impurity**: A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. **index** and **entropy** are two commonly used impurity measurements in decision trees for classifications task
- **Variance**: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. **Mean squared error, Mean Absolute Error, friedman_mse, or Half Poisson deviance** are used to measure the variance for the regression tasks in the decision tree.
- **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
- **Pruning**: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.

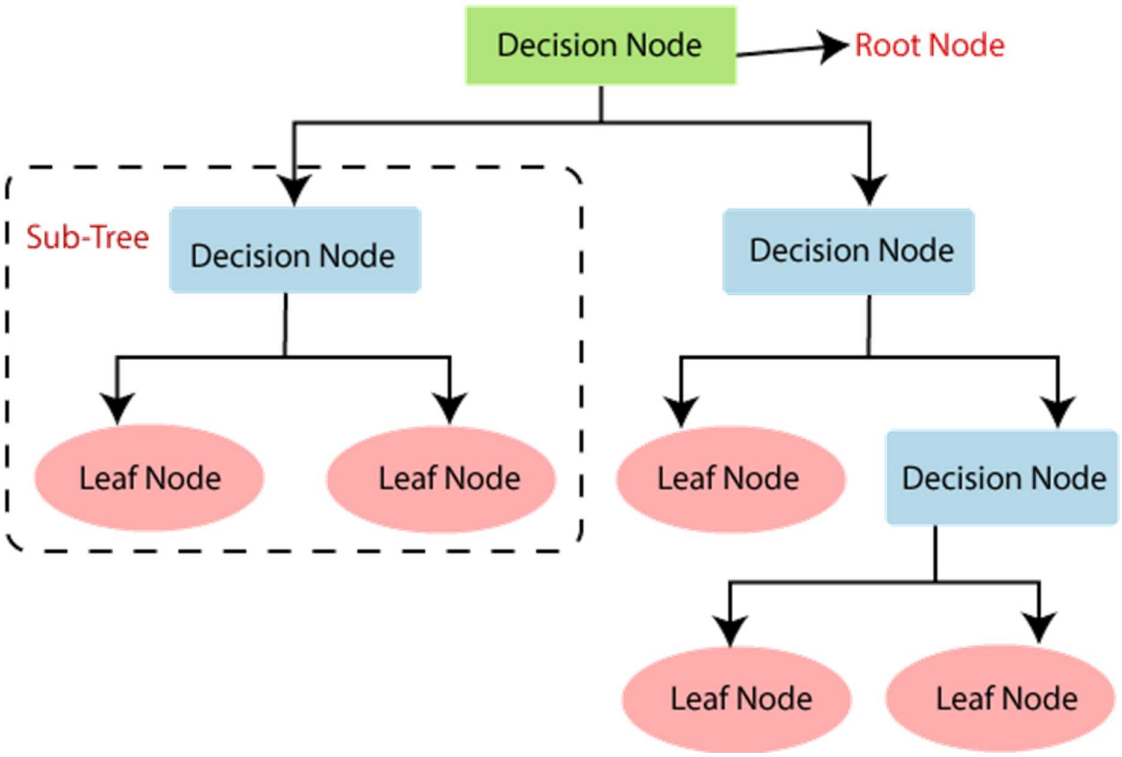

Fig.4.6:decision tree

3. Support Vector Machine:

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data. This is done by finding the hyperplane that has the largest margin, which is defined as the distance between the hyperplane and the closest data points from each class. Once the hyperplane is determined, new data can be classified by determining on which side of the hyperplane it falls. SVMs are particularly useful when the data has many features, and/or when there is a clear margin of separation in the data.

What are Support Vector Machines? Support Vector Machine (SVM) is a relatively simple **Supervised Machine Learning Algorithm** used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to use for multi-class problems. **Support Vector Machine for Multi-CLass Problems** To perform SVM on multi-class

problems, we can create a binary classifier for each class of the data. The two results of each classifier will be :

- The data point belongs to that class OR
- The data point does not belong to that class.

For example, in a class of fruits, to perform multi-class classification, we can create a binary classifier for each fruit. For say, the 'mango' class, there will be a binary classifier to predict if it IS a mango OR it is NOT a mango. The classifier with the highest score is chosen as the output of the SVM. **SVM for complex (Non Linearly Separable)** SVM works very well without any modifications for linearly separable data. **Linearly Separable Data** is any data that can be plotted in a graph and can be separated into classes using a straight line.

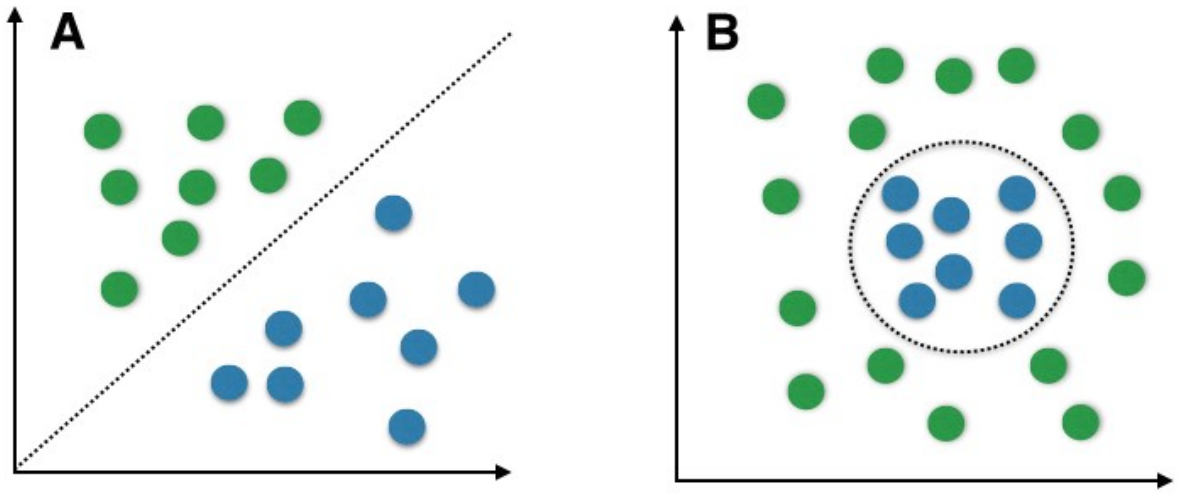

Fig.4.7:SVM

We use **Kernelized SVM** for non-linearly separable data. Say, we have some non-linearly separable data in one dimension. We can transform this data into two dimensions and the data will become linearly separable in two dimensions. This is done by mapping each 1-D data point to a corresponding 2-D ordered pair. So for any non-linearly separable data in any dimension, we can just map the data to a higher dimension and then make it linearly separable. This is a very powerful and general transformation. A **kernel** is nothing but a measure of similarity between data points. The **kernel function** in a kernelized SVM tells you, that given two data points in the original feature space, what the similarity is between the points in the newly transformed feature space. There are various kernel functions available, but two are very popular :

- **Radial Basis Function Kernel (RBF):** The similarity between two points in the transformed feature space is an exponentially decaying function of the distance between the vectors and the original input space as shown below. RBF is the default kernel used in SVM.

$$K(x, x') = exp(-\gamma||x - x'||)$$

- **Polynomial Kernel:** The Polynomial kernel takes an additional parameter, 'degree' that controls the model's complexity and computational cost of the transformation

A very interesting fact is that SVM does not actually have to perform this actual transformation on the data points to the new high dimensional feature space. This is called the **kernel trick**. **The Kernel Trick:** Internally, the kernelized SVM can compute these complex transformations just in terms of similarity calculations between pairs of points in the higher dimensional feature space where the transformed feature representation is implicit. This similarity function, which is mathematically a kind of complex dot product is actually the kernel of a kernelized SVM. This makes it practical to apply SVM when the underlying feature space is complex or even infinite-dimensional. The kernel trick itself is quite complex and is beyond the scope of this article. **Important Parameters in Kernelized SVC (Support Vector Classifier)**

1. **The Kernel**: The kernel, is selected based on the type of data and also the type oftransformation. By default, the kernel is Radial Basis Function Kernel (RBF).
2. **Gamma** : This parameter decides how far the influence of a single training example reaches during transformation, which in turn affects how tightly the decision boundaries end up surrounding points in the input space. If there is a small value of gamma, points farther apart are considered similar. So more points are grouped together and have smoother decision boundaries (maybe less accurate). Larger values of gamma cause points to be closer together (may cause overfitting).
3. **The 'C' parameter**: This parameter controls the amount of regularization applied to the data. Large values of C mean low regularization which in turn causes the training data to fit very well (may cause overfitting). Lower values of C mean higher regularization which causes the model to be more tolerant of errors (may lead to lower accuracy.

CHAPTER-V

EXPECTED RESULTS AND DISCUSSION

Chapter – V:

EXPECTED RESULTS AND DISCUSSION

To evaluate the performance of the proposed model, various performance measures are employed, including accuracy, precision, F1 score, and recall. These metrics are crucial in assessing the model's ability to accurately classify diseases in tomato plants. The equations (1-4) below illustrate how these performance measures are calculated [18]. In these equations, "True Positive" is represented as "TP," "True Negative" as "TN," "False Positive" as "FP," and "False Negative" as "FN":

- $\boldsymbol{Precision} = \boldsymbol{TP}\ \boldsymbol{TP}+\boldsymbol{FP}$
- $\boldsymbol{Recall} = \boldsymbol{TP}\ \boldsymbol{TP}+\boldsymbol{FN}$
- $\boldsymbol{Accuracy} = (\boldsymbol{TP}+\boldsymbol{TN})\ [(\boldsymbol{TP}+\boldsymbol{FP})+(\boldsymbol{TN}+\boldsymbol{FN})]$
- $\boldsymbol{Fl\ score} = \mathbf{2}\ \boldsymbol{Precision}*\boldsymbol{Recall}\ (\boldsymbol{Precision}+\boldsymbol{Recall})$

By calculating these performance measures, we gain insights into the effectiveness of the model and its ability to understand better in customer data.

Below are the some of data visualizations of amazon dataset after preprocessing.

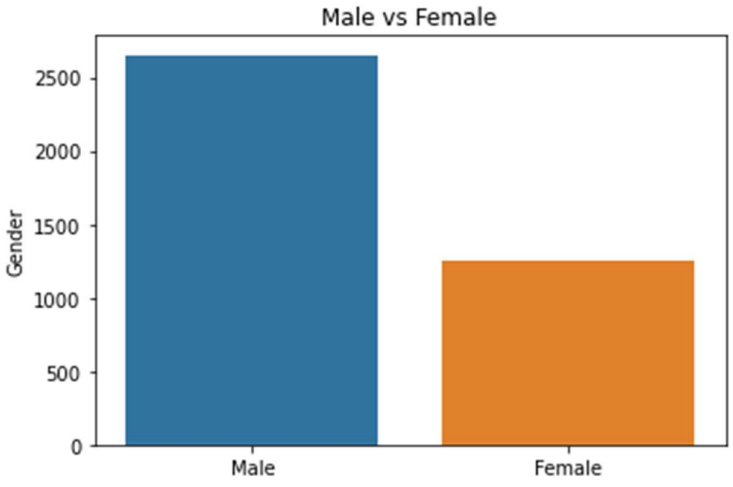

Fig.5.1:Visualization according to gender

Implementation of dataset according to different types of payment modes.

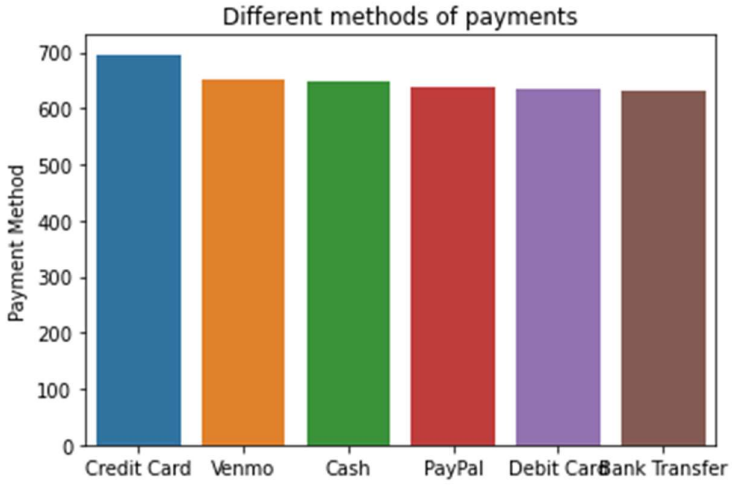

Fig.5.2:visualization according to payment mode

Datavisualization of amazon dataset according to sizes preferred.

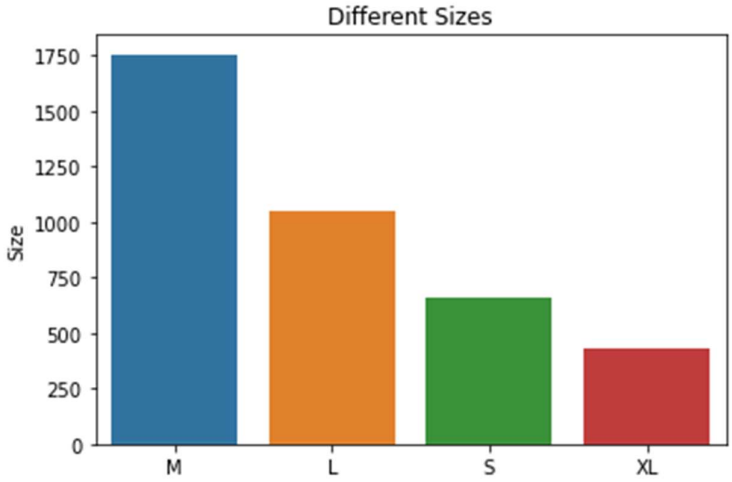

Fig.5.3:visualization according to sizes

Now let us discuss about the different types of machine learning techniques used such as SVM,KNN,Decision tree,Random forest,Naive Bayes.

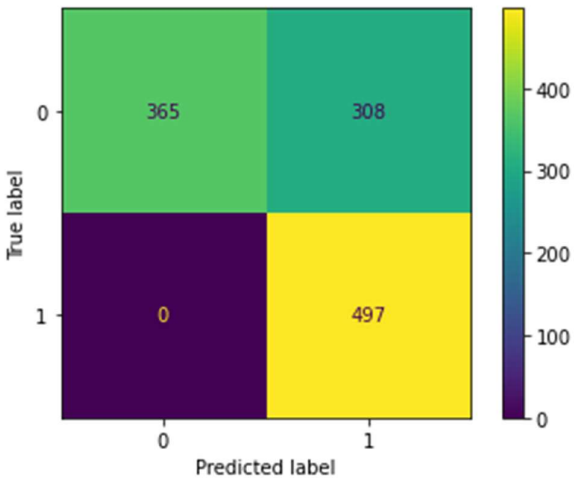

Fig.5.4:Random Forest

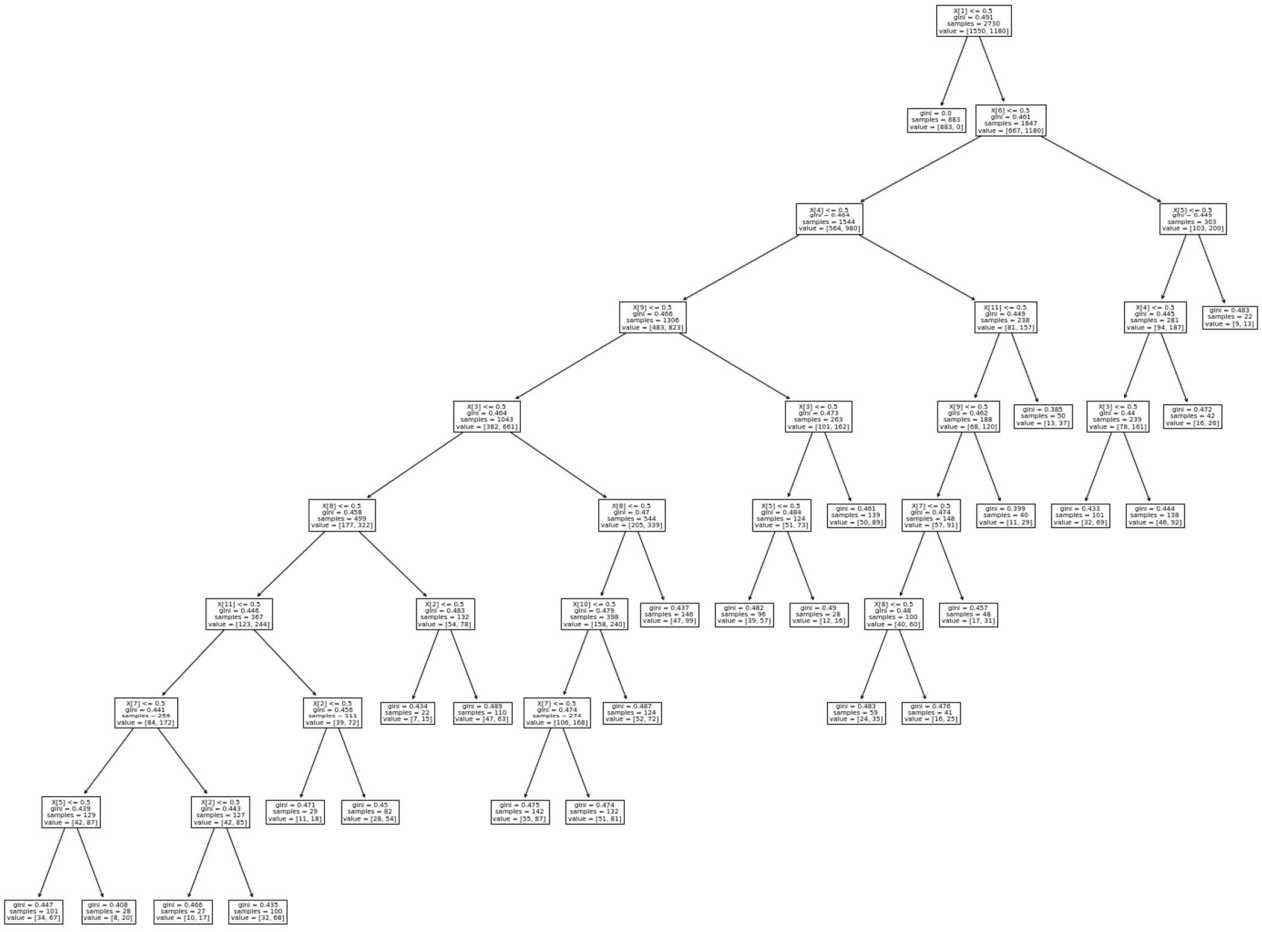

Fig.5.5:Decision tree

CHAPTER-VI
CONCLUSION AND FUTURE SCOPE

Chapter – VI:

CONCLUSION AND FUTURE SCOPE

In conclusion,Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining can help improve customer engagement by:

- Understanding customer preferences, needs, and behavior patterns
- Segmenting customers into groups based on their similarities and differences
- Personalizing products, services, and offers to match customer expectations
- Predicting customer outcomes, such as churn, retention, loyalty, and satisfaction
- Recommending products, services, or actions that are relevant and beneficial to customers
- Improving customer service and support by identifying and resolving issues
- Enhancing customer communication and feedback by using the appropriate channels and methods

Data mining can be performed using various techniques, such as anomaly detection, association rule learning, classification, clustering, and regression analysis. These techniques can help extract valuable insights from different types of data, such as text, images, audio, video, and social media. Data mining can also be integrated with other technologies, such as artificial intelligence, machine learning, and natural language processing, to create more advanced and intelligent solutions for customer engagement.
Data mining can provide a competitive advantage for businesses that want to attract, retain, and delight their customers. By using data mining, businesses can create a better customer experience, increase customer loyalty, and generate more revenue. However, data mining also poses some challenges, such as data quality, privacy, security, and ethics.

Futurescope:
Using this Data Mining techniques we can make a Recommendation syastems for different types of ecommerce websites.A recommendation system is a system that suggests items or actions to users based on their preferences, needs, or behavior. Data mining is the process of analyzing large amounts of data to discover patterns, trends, or insights. Data mining techniques can be used to build a recommendation system by:

- Finding the similarity or dissimilarity between users or items based on their features or ratings
- Segmenting users or items into groups or clusters based on their similarity or dissimilarity
- Predicting the ratings or preferences of users for items based on their past ratings or preferences
- Generating rules or associations that indicate the co-occurrence or correlation of items or actions
- Personalizing the recommendations based on the user’s profile, context, or feedback

Some of the common data mining techniques that are used for recommendation systems are:

- Collaborative filtering: This technique uses the ratings or feedback of other users who have similar tastes or preferences to recommend items to a user. For example, if user A and user B both liked items X and Y, then item Z, which is liked by user B, can be recommended to user A.
- Content-based filtering: This technique uses the features or attributes of the items to recommend items to a user based on their similarity or relevance to the user's preferences or needs.
- Hybrid filtering: This technique combines collaborative filtering and content-based filtering to overcome the limitations or drawbacks of each technique. For example, collaborative filtering may suffer from cold start problem, which occurs when there is not enough data about new users or items, while content-based filtering may suffer from overspecialization problem, which occurs when the recommendations are too narrow or similar to the user's preferences.
- Association rule mining: This technique uses the frequency or co-occurrence of items or actions to generate rules that indicate the relationship or correlation between them. For example, if many users buy bread and butter together, then a rule can be generated that says "if bread, then butter". These rules can be used to recommend items or actions to users based on their previous or current choices.
- Community detection: This technique uses the network or graph structure of the data to find groups or communities of users or items that are densely connected or related to each other. For example, if users are connected by friendship links and items are connected by similarity links, then communities of users or items can be found that share common interests or preferences. These communities can be used to recommend items or actions to users based on their membership or affiliation.

CHAPTER – VII

REFERENCES

Chapter- VII:

REFERENCES

[1]GeeksForGeeks SVM

[2] Analyticsvidhya Random Forest

[3] Data Mining

[4] Scikit-learn tree

[5] IBM KNN

[6] Simplilearn Naïve Baye

[7] Kaggle Amazon Sales Dataset

[8] Kaggle Customer Shopping Trends Dataset

[9] Kaggle Amazon Sales Reports