UNIT-III

Mining Frequent patterns :-

Frequent pattern mining in data mining is the process of identifying patterns or associations within a dataset that occur frequently. This is typically done by analyzing large datasets to find items or sets of items that appear together frequently.

* Frequent pattern extraction is an essential mission in data mining that intends to uncover repetitive patterns or itemsets in a granted dataset.
* It encompasses recognizing collections of components that occur together frequently in a transactional or relational database.
* This procedure can offer valuable perceptions into the connections and affiliations among diverse components or features within the data.

Algorithms used for frequent pattern mining:

1. Aprion algorithm:

This is one of the most commonly used algorithms for frequent pattern mining. It uses a "bottom-up" approach to identify

frequent itemsets and then generates association rules from thase itemsets.

2. ECLAT algorithm:-
This algorithm uses a "depth-first search" approach to identify frequent itemsets. It is particularly efficient for datasets with a large number of items.

3. FP-growth algorithm:
This algorithm uses a "compression" technique to find frequent patterns efficiently. It is particularly efficient for databases with a large number of transactions.

Applications:-

* Market basket analysis
* customer behaviour analysis
* web mining
* bioinformatics
* network traffic analysis

Advantages:-

1. It can find useful information which is not visible in simple data browsing.
2. It can find interesting association and correlation among data items.

Disadvantages:-

1. It can generate a large number of patterns.

2. With high dimensionality, the number of pattems can be very large, making it difficult to interpret the results.

Associations and Correlations:-

Association :-

* Association deals with uncovering patterns or Co-occurances between items, such as which products are frequently purchased together in market basket analysis.
* Association rules take the form "if x then y" to indicate the tendency of certain items or events to occur together.
* Algorithms like Aprion and FP-Growth are used to mine frequent itemsets and generate association rules from large transaction datasets.

Applications:-

1) Market basket analysis
2) Healthcare
3) Cybersecurity
4) inventory management

Advantages:-

1) Helps find useful patterns that can guide business choices.
2) Works well with lots of data and is easy for people to understand the results.

Disadvantages:-

* Can create too many rules, some of which are not important or useful.
* Needs clean, accurate data; otherwise the results may be misleading.

Correlation:-

* Correlation refers to the statistical relationship or dependency between items or itemsets that occur together frequently in the data.
* It is used to measure how strongly the occurance of another, beyond just their frequent co-occurance.
* Correlation measures how much the presence of one itemset influences the presence of another itemset, often evaluated beyond simple support and confidence metrics.
* A key measure often used for correlation in frequent pattern is lift, which compares the observed co-occurrence of items with the expected co-occurrence if they were independent.
* other correlation measures includes all confidence, kulczynski and cosine similarity, which provide different perspectives on how itemsets relate based on frequency and co-occurance.

Advantages:-

1) Helps focus on strong and useful connections between items.

2) Filters out random or weak relationships, so the results are more useful.

Disadvantages:-

1) Can be a bit harder to calculate than just counting how often items occur.

2) Needs good quality data for accurate results.

3) A high correlation does not mean one item causes the other, just that they are related.

Mining methods:-

There are many methods used for Data mining, but the crucial step is to select the appropriate form from them according to the business or the problem statement. These methods help in predicting the future and then making decisions accordingly. These also help in analyzing market trends and increasing company revenue.

Some methods are:

1) Association
2) Classification
3) Clustering Analysis
4) Prediction
5) Sequential patterns or Pattern Tracking

6) Decision trees

7) Outlier Analysis or Anomaly Analysis

8) Neural Network

1. Association:-

It is used to find a correlation between two or more items by identifying the hidden pattern in the data set and hence also called relation analysis. This method is used in market basket analysis to predict the behaviour of the customer.

There are two types of Association Rules:

1) Single dimensional association rule: These rules contain a single attribute that is repeated.

2) Multi dimensional association rule: These rules contain multiple attributes that are repeated.

2. Classification:-

This data mining method is used to distinguish the items in the data sets into classes or groups. It helps to predict the behaviour of entities within the group accurately.

It is a two-step process:

* Learning step: In this, a classification algorithm builds the classifier by analyzing a training set.
* Classification step: Test data are used to estimate the accuracy or precision of the classification rules.

3. Clustering Analysis:-

Clustering is almost similar to classification, but in this cluster are made depending on the similarities of data items. Different groups have dissimilar or unrelated objects. It is also called data segmentation as it partitions huge data sets into groups according to the similarities.

Various clustering methods are used:

1) Hierarchical Agglomerative methods
2) Grid based methods
3) partitioning methods
4) Model-based methods
5) Density-Based methods

4. prediction:-

This method is used to predict the future based on the past and present trends or data set. Prediction is mostly used to combine other mining methods such as classification, pattern matching, trend analysis, and relation.

Regression analysis is the best choice to perform prediction. It can be used to set a relationship between independent variables and dependent variables.

5. Sequential patterns or pattern tracking

This method is used to identify patterns that frequently occur over a certain period of time.

For example, a clothing company's sales manager sees that sales of jackets seem to increase just before the winter season, or sales in bakery increase during Christmas or New years eve.

6. Decision trees:-

A decision tree is a tree structure, where

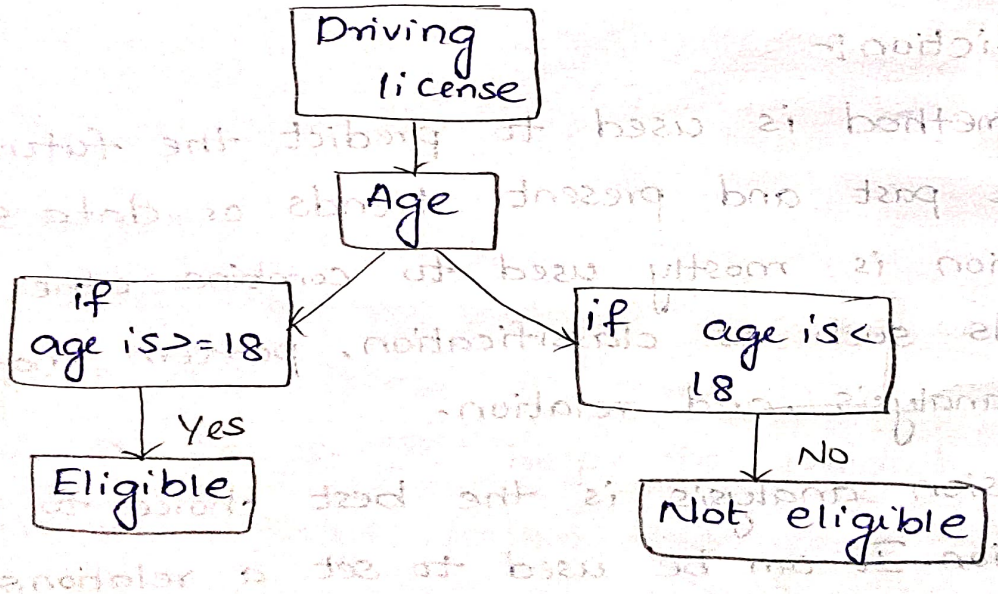
→ Each internal node represents a test on attribute.

→ Branch denotes the result of the test.

→ Teminal nodes holds the class label.

→ The topmost node is the root node which has a simple question that has two or more answers.

Example:-



7. Outlier Analysis or Anomaly Analysis:-

This method identifies the data items that do not comply with the excepted pattern or excepted behaviour. These unexcepted data items are considered as outlier or noise. They are helpful in many domains like credit card

fraud detection, intrusion detection, faullt detection etc. This is also called outlier mining.

8. Neural Network:-

This method or model is based on biological neural networks. It is a collection of neurons like processing units with weighted connections between them. They are used to model the relationship between inputs and outputs. It is used for classification, regression analysis, data processing etc. This technique works on three pillars-

1) Model
2) Learning algorithm
3) Activation function

Pattern evaluation method:-

Pattern evaluation in data mining is the process of assessing the quality, usefulness, and importance of discovered patterns, such as association rules or sequential patterns. It involves using various objective and subjective measures to determine the validity, strength, and relevance of the patterns to ensure they are statistically significant and practically useful.

purpose :-

1) To remove redundant or irrelevent patterns.
2) To find truly valuable insights that can help

in decision-making.

3) To measure the interestingness or usefulness of discovered knowledge.

Pattern evaluation methods :-

1. objective Measures:-

Support: Frequency of the pattern in the dataset; important for determining how common a pattern is.

Confidence: Reliability of an association rule; probability that the consequent occurs given the antecedent.

Lift: Strength of the association by comparing observed support to expected support if items were independent; a lift > 1 indicates a meaningful pattern.

2. Subjective Measures:-

User Interest: Pattern relevence based on user needs or domain goals.

Surprisingness: Whether the pattern reveals unexpected or new insights.

Novelty: New or previously unknown patterns are often more interested.

3. Statistical Significance Tests :-

chi-square Test: Measures independence of variables in the patterns.

p-value: Probability that a pattern is due to chance; low p-values indicate significance.

4. Information-Theoretic Measures:-

Entropy: Measures uncertainity; lower entropy means more informative

Mutual Information: Amount of shared information between variables.

Pattern mining in multilevel, multidimensional space:-

Pattern mining in multilevel and multidimensional space refers to discovering frequent patterns or associations considering data organized at various abstraction levels and involving multiple attributes or dimensions.

Multilevel pattern mining:-

* Involves mining frequent itemsets or association rules across different levels of abstraction within a concept hierarchy.
* A concept hierarchy organizes data from general to specific.
* The mining typically follows a top-down approach: starting at a high abstraction level to find frequent patterns, then drilling down to lower levels to discover more detailed pattern.
* This method helps identify meaningful insights at different granularities and avoids common-sense patterns found only at high-levels.

* Algorithms like Apriori are adapted to handle multilevel associations with varying support thersholds for different.

Multidimensional pattern mining:-

* Involves mining association rules that span multiple dimensions or attributes of the data, such as customer purchase behaviour linked at age, occupation, or location.
* Patterns include relationships between predictates across dimensions.
* Data cubes are often used to represent aggregates in multidimensional space for efficient mining.
* Both categorical and quantitative attributes are handled, with approaches like discretization, clustering, and gradient methods for numerical data.

Advantages:-

1) Improved knowledge Discovery
2) Enhanced Flexibility
3) Multi-Attribute relationships
4) Scalability and efficiency
5) discovery of rare and negative patterns
6) Business and scientific impact

disadvantages:-

1. Increased complexity
2. Data preparation challenges
3. Risk of overfitting and Redudancy
4. Handling sparse and noisy data
5. scalability limitations

Constraint based Frequent pattern mining:-

A data mining process may uncover thousands of rules from a given data set, most of which end up being unrelated or uninterstting to users. Often, users have a good sense of which "direction" of mining may lead to interesting patterns and the "form" of the patterns or rules they want to find. They may also have a sense of "conditions" for the rules, which would eliminate the discovery of certain rules that they know would not be of interest. Thus, a good heuristic is to have the users specify such intuition or expectations as constraints to confine the search space. This strategy is known as constraint-based mining.

The constraints can include the following:

knowledge type constraints: These specify the type of knowledge to be mind, such as association, correlation, classification or clustering.

Data constraints: These specify the set of task-relevant data.

Dimension / level constraints: These specify the desired dimensions of the data, the abstraction levels, or the level of the concept hierarchies to be used in mining.

Interestingness constraints:- These specify thersholds on statistical measures of rule interestingness, such as support, confidence and correlation.

Rule constraints:- These specify the form of rules to be mined. Such constraints may be expressed as metarules, as the maximum or minimum number of predicates that can occur in the rule antecedent or consequent, or as relationships among attributes, attribute values, and/or aggregates.

Classification using Frequent patterns:-

Classification using frequent patterns is a data mining technique that builds predictive models by identifying sets of items or features (called frequent patterns or itemsets) that commonly appear together in a dataset and then uses these patterns to classify new data points.

* The process begins by mining frequent patterns from the dataset using algorithms such as Apriori, FP- Growth, or ECLAT, which look for combinations of attributes or items that appear above a minimum support threshold.
* Once these patterns are identified, they are turned into classification rules. Each rule connects a frequent pattern with a particular class label, usually in an "if-then" form.
* Not all mined patterns are equally good for classification - the best score are chosen using metrics like confidence the pattern predicts a, and the final classifier uses these rules to predict new instances.
* This method not only achieves high accuracy in prediction but also provides insights into relationships between different features in the data.

Methods and algorithm:-

The algorithms used are

1) Apriori algorithm
2) FP- Growth algorithm
3) closed frequent itemset mining
4) Naive Bayesian algorithm

1. Apriori algorithm:-

The apriori algorithm is an algorithm for finding frequent item sets in a given dataset. It is an unsupervised learning technique that employs a "bottom-up" strategy to discover frequent itemsets in a dataset by first recognizing indivisual items in the dataset and then looking for combinations of items that appear often together. The Apriori technique may be used to identify the rules that govern relationships between various objects in a collection. It is frequently used in market basket analysis.

2. FP-Growth algorithm:-

The FP-Growth algorithm is a data mining technique that finds frequent patterns or itemsets in a dataset. It operates by building an FP-Tree, which is a compact representation of the dataset. The FP-Tree is then utilized to construct common patterns from the ground up. The FP-Growth technique is very scalable and can efficiently detect common patterns in huge datasets.

3. Closed frequent itemset mining:-

* Closed frequent itemset mining is a kind of frequent itemset mining in which all itemsets

in a given dataset with a frequency that meets or exceeds a predetermined thershold are discovered.

* The technique works by first generating a list of all frequent itemsets in the dataset, then iteratively evaluating each item set to determine whether any supersets of itemsets have a frequency that meets thershold. Any supersets that meet the criteria are added to the list of frequently occuring itemsets. This procedure is continued until no further supersets are discovered.

4. Naive Bayeisan Algorithm:-

* Naive Bayes is a form of supervised machine learning method that is used for classification and is based on Baye's theorem. It is a probabilistic method that predicts using the probability of each attribute belonging to each class. The Naive Bayes method is based on assumption that all qualities are independent on another.

* This streamlines the computation of probabilities, allowing the algorithm to easily forecast a class of new data point.

Examples and Applications:-

1) Market basket analysis
2) Image classification
3) Healthcare and recurement
4) Fraud detection.

Advantages:-

1) Easily finds significant combinations of features that help in classification, rather than relying on indivisual feature ones.

2) Creates clear and interpretable "if-then" rules, making the classification rules easy to understand by humans.

3) Improves prediction accuracy by using the relationships between features.