**Project 10: Market Basket Insights**

Data visualization is the process of representing data and information graphically through charts, graphs, maps, and other visual elements. Its primary goal is to make complex datasets more understandable, accessible, and interpretable by presenting them in a visual format. Data visualization leverages the human visual system's ability to quickly recognize patterns, trends, and anomalies, allowing users to gain insights from data more effectively than they would from raw numbers or text.

**Key aspects of data visualization include:**

1. **Data Representation:** Data is transformed into visual elements such as bars, lines, points, or shapes, which are organized in a way that conveys meaningful information.
2. **Visual Encoding:** Attributes like position, length, color, size, and shape are used to represent data values. For example, a bar chart may use the length of bars to represent numeric values.
3. **Interactivity:** Many modern data visualizations are interactive, allowing users to explore data by hovering over elements, zooming, filtering, or selecting specific data points to reveal more details.
4. **Storytelling:** Data visualizations can be used to tell a story or convey a message by carefully selecting and arranging visual elements to emphasize specific insights or trends.
5. **Exploration and Analysis:** Data visualization tools and techniques help analysts and decision-makers explore data, discover patterns, outliers, and trends, and draw conclusions from the information presented.

Today, we have access to a plethora of data visualization tools and software that empower individuals and organizations to create stunning and informative visuals. Some of the popular options include:

* **Tableau**: Known for its user-friendly interface and robust features, Tableau is a favorite among data professionals.
* **Power BI**: Microsoft's Power BI is another strong contender, offering seamless integration with other Microsoft products.
* **D3.js**: For those who want complete control and customization, D3.js is a JavaScript library that lets you build interactive visualizations from scratch.
* **Python Libraries**: Python enthusiasts can leverage these libraries to create data visualizations within their coding environment.

**Implementation:**

This project is implemented in java. For the user interface to provide the input data java swing is used to design the interface. The apriori algorithm is used to process the data and generated the association rule as a output in a file.

**Tools used:**

**Intellij**: Intellij was used as a IDE to develop this application

**Java swing**: Java swing was used for designing the user interface.

**Java**: java programming language was used to implement the algorithm

**File**: File is used as the database to process the data.

**Description of major classes and methods:**

Main Window

This class is used to display the user interface for providing the input to the system. This class extends the JFrame class Some of the important methods of this class are:

ActionPerformed()

askUserToChooseInputFile()

askUserToChooseOutputFile()

AlgoApriori

This is the main class that executes the algorithm after user presses the run algorithm button.This class performes the processiong of the data from user input and display output to the user decided output file.

Some of the important methods of this class are:

runAlgorithm(double minsup, String input, String output)

generateCandidate2(List frequent1)

generateCandidateSizeK(List levelK\_1)

printStats()

Itemset

Itemset This class stores the items as the array of integer and provides the support of the respective item from the given input data.

Some of the important methods of this class are:

getAbsoluteSupport()

getItems()

**Testing:**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. Unit testing was performed to test correctness of different modules.

**Testing correctness of the output:**

In this test approach sample input was given to the system and different support and confidence was provided. At first large sample input were given to the system with different support and confidence value. Then few sample input were given to the system with different support and confidence.

Table 5- Test result for large sample input

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No of transaction | Support | Confidence | No of rules generated | No rules with confidence |
|  | **10%** | **40%** | **21** | **12** |
| **24** | **20%** | **40%** | **2** | **1** |
|  | **30%** | **40%** | **0** | **0** |

As show in the table 5 sample data of 24 transaction was taken as input to the system and when the support was 10% and confidence 40% 21 rules were generate from which 12 rules were found to be strong with 100% confidence. When the support was 20% with same confidence then 2 rules was generated from which 1 rule was found o be strong and when support was 30 % no rule was generated.

Table 6- Test result for small sample input

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No of transaction | Support | Confidence | No of rules generated | No rules with confidence=100% |
|  | **10%** | **50%** | **62** | **31** |
| **7** | **10%** | **60%** | **31** | **101** |
|  | **20%** | **70%** | **10** | **10** |
|  | **30%** | **70%** | **0** | **0** |

As show in the Table 6 sample data of 7 transaction was taken as input to the system and when the support was 10% and confidence 50% 62 rules was generated from which 31 rules was found to be strong with 100% confidence. When the support was 10% with 60% confidence then 31 rules was generated from which 31 rules was found to be strong. When support was 20% and confidence was 70% 10 rules was generated from which 10 rules was found to be strong and when support was 30 % no rule was generated.