# AIN429 Data Mining Laboratory

**Assignment 3: Frequent Pattern Mining**

**Due Date**: 27/11/2024 23:59

## Aim of the Experiment

In this assignment, you will explore frequent pattern mining techniques, focusing on mining frequent itemsets using the Apriori and FP-Growth algorithms. These are analytical processes that find frequent patterns, associations, or causal structures from data sets. You will implement both algorithms and analyze their performance on a real-world dataset. The assignment should be implemented as a single Jupyter Notebook. Use comments and Markdown cells to document your work, and ensure reproducibility of your analysis.

## Part 1: Dataset Import, Preprocessing, and Analysis (15 Marks)

**Tasks:**

1. Load the provided dataset and print basic information, such as the number of transactions, unique items, and any missing values etc.
2. Report key statistics about the dataset (e.g., average number of items per transaction, frequency of items etc.).
3. Clean and transform the dataset into a format suitable for frequent pattern mining algorithms (e.g., transaction lists).
4. Remove any infrequent items or handle missing data appropriately.
5. Create visualizations to represent item frequencies and transaction lengths.

## Part 2: Apriori Algorithm Implementation and Application (35 Marks)

**Tasks:**

1. **Implementation**:

   - Implement the Apriori algorithm to find frequent itemsets in the dataset.
   - You will be implementing the algorithm yourselves, the following are the main steps of the algorithm:
     i. Calculate the support of item sets (of size k = 1) in the transactional database (note that support is the frequency of occurrence of an itemset). This is called generating the candidate set.
     ii. Prune the candidate set by eliminating items with a support less than the given threshold.
     iii. Join the frequent itemsets to form sets of size k + 1, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support

- Use a minimum support threshold of 0.01.
- Try the algorithm with different threshold values.

2. **Results**: Display the frequent itemsets found with their respective support values.

3. **Analysis**:

- Summarize and interpret the frequent patterns discovered using Apriori.
- Discuss the computational complexity and runtime for different support thresholds.

---

## Part 3: FP-Growth Algorithm Implementation and Application (35 Marks)

**Tasks:**

1. **Implementation**:

- Implement the FP-Growth algorithm to find frequent itemsets in the dataset, again by yourselves.
- The following are the main steps of the algorithm:
    - i. Calculate the support of individual items in the transactional database and remove items below the support threshold.
    - ii. Sort the remaining items in descending order of support and rearrange transactions accordingly.
    - iii. Construct an FP-Tree, grouping transactions with shared prefixes while maintaining support counts for items.
    - iv. For each item in the FP-Tree, find all paths where the item appears and build a smaller conditional FP-Tree for those paths.
    - v. Recursively mine the conditional FP-Trees to extract all frequent itemsets.
- Use the same minimum support threshold (0.01).
- Try the algorithm with different threshold values.

2. **Results**: Display the frequent itemsets found with their respective support values.

3. **Analysis**:

- Summarize and interpret the frequent patterns discovered using FP-Growth.
- Discuss the computational complexity and runtime for different support thresholds.

## Part 4: Visualization and Comparison (15 Marks)

**Tasks:**

1. **Performance Comparison**:
    - Compare the performance of Apriori and FP-Growth algorithms
    - Use tables and graphs to visualize the comparison.
2. **Insights**:
    - Summarize the key findings and patterns from the dataset.
    - Discuss the advantages and limitations of both algorithms.

## Report

Write concise explanations (in markdown cells) summarizing your findings. The report should be done alongside the previous tasks and should include:

1. The dataset and preprocessing steps.
2. The frequent patterns discovered and their potential implications.
3. The performance comparison results.

## Grading Breakdown

- **Dataset Import, Preprocessing, and Analysis**: 15%
- **Apriori Algorithm**: 35%
- **FP-Growth Algorithm**: 35%
- **Visualization and Comparison**: 15%

## Submission Requirements

- **Format**: Submit your assignment as b`<studentID.zip>` containing:
    - A single Jupyter Notebook (`assignment_3.ipynb`).
- **Late Penalty**: 10% per day, up to 2 days.
- **Submission Method**: Use https://submit.cs.hacettepe.edu.tr.
    - **Not getting 1 from the submit system will lose you points!**
- **Important** The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- **Questions** You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.

## Helpful Links

- https://towardsdatascience.com/frequent-pattern-mining-association-and-correlations-8fa9f80c22ef
- https://www.geeksforgeeks.org/apriori-algorithm/
- https://www.geeksforgeeks.org/ml-frequent-pattern-growth-algorithm/2
- http://rasbt.github.io/mlxtend/user_guide/preprocessing/TransactionEncoder/