Data Mining Assignment 1

Identify a problem from your own experience that you think would be amenable to data mining. For that problem describe:

1. What the data is.  
2. What type of benefit you might hope to get from data mining.  
3. What type of data mining (classification, clustering, etc.) you think would be relevant.  
4. Name one type of data mining that you think would not be relevant, and describe briefly why not.  
For each, illustrate with an example, e.g., if you think clustering is relevant, describe what you think a likely cluster might contain and what the real-world meaning would be.

Write one to two pages of 11 point single-spaced typeset text - you aren't writing a paper, but it isn't short answer either.

**1. What the data is?**

Ans) Data is the information that has been translated into a form that is efficient for movement or processing. This information can be in the form of audio clips, documents, images and programs. Based on the requirement of the problem, the data can be modified into the required format.

**2. What type of benefit you might hope to get from data mining.**

Data mining involves collecting, processing and analyzing the data to discover the insights from it. There are various techniques and methodologies involves and serves different purposes. Data mining helps organizations in analyzing the huge amount of data and find out the new facts according to the goal for which you are using it. Some of the techniques are as follows

**Clustering:** Use to merge and classify the data under various categories which are helpful in future analysis.

**Regression:** It is a technique which aims to predict the future behaviour of collected large sets of data.

**Anomaly detection:** This technique helpful in identifying the abnormalities in the existing data sets.

**3. What type of data mining (classification, clustering, etc.) you think would be relevant.**

Classification and clustering are two methods of **pattern identification used in machine learning.** Although both techniques have certain similarities, the difference lies in the fact that classification uses **predefined classes** in which objects are assigned, while clustering **identifies similarities between objects**. Clustering is used in projects for companies that want to find **common aspects within their customers** to find groups and focus products or services. Whereas, Classification is used when you need to **know users or customers** to decide which products or campaigns will be launched in the future. So, in my view clustering is more relevant than the classification algorithm.

**4. Name one type of data mining that you think would not be relevant, and describe briefly why not.  
For each, illustrate with an example, e.g., if you think clustering is relevant, describe what you think a likely cluster** **might contain and what the real-world meaning would be.**

I think classification is not relevant for the problem I considered. Classification is used where I use the trained data set to classify the values I have.  
In this problem, if I need to follow classification, based on the previous products a customer bought, I should be able to say whether he’ll buy a product I have or not.

For suppose, a customer ‘x’ bought milk and jam. Then, I can say whether he would buy the bread or not in classification.

So, I think classification is not relevant for this problem.