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?

Data mining is the process that companies use to turn raw data into useful information. They utilize software to look for patterns in large batches of data so they can learn more about customers. It pulls out information from data sets and compares it to help the business make decisions. This eventually helps them to develop strategies, increase sales, market effectively.

Business understanding. The first step to successful data mining is to understand the overall objectives of the business. For example, a supermarket may want to use data mining to learn more about their customers. The business understanding is that a supermarket is looking to find out what their customers are buying the most.

Data understanding. After you know what the business is looking for, it's time to collect data. There are many complex ways that data can be obtained from an organization, organized, stored, and managed. Data mining involves getting familiar with the data, identifying any issues, getting insights, or observing subsets. For example, the supermarket may use a rewards program where customers can input their phone number when they purchase, giving the supermarket access to their shopping data.

Data Preparation. Data preparation involves getting the information production ready. This is the biggest part of data mining. It is taking the computer-language data, and converting it into a form that people can understand and quantify. Transforming and cleaning the data for modeling is key for this step.

Modeling. In the modeling phase, mathematical models are used to search for patterns in the data. There are usually several techniques that can be used for the same set of data. There is a lot of trial and error involved in modeling.

Evaluation. When the model is complete, it needs to be carefully evaluated and the steps to make the model need to be reviewed, to ensure it meets the business objectives. At the end of this phase, a decision about the data mining results will be made. In the supermarket example, the data mining results will provide a list of what the customer has purchased, which is what the business was looking for.

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.