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?

Ans: The data contains the customer history related to loan and credit rating. This includes, employment tracking, credit history, home/rent history, type of investments, etc.  
2. What type of benefit you might hope to get from data mining?

Ans: Data mining here helps in predicting the customer viable for loan applications and also predicts probability of each customer credit ratings.  
3. What type of data mining (classification, clustering, etc.) you think would be relevant?

Ans: Classification assigns items from a collection to target classes. Hence, classification is used in this type of application and helps in predicting the loan applicants who are preferable.  
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.

Ans: We are using classification data mining technique as here we need pre-defined target classes to categorize the new data. Classification has no order and is discrete. Hence there can be different classifications possible like binary classifications with 2 values as output or multiple classifications according to the requirement.

Here we have to check whether an individual is fit for loan application or not. For this the data provided contains the history that states his credit rate, employment track, home or rent history, his number and type of investments, etc. These are the different predictors used to classify the individual to the target. As part of the training process, we find the relationships between values of the predictors and the values of the target. This helps us train the model. As part of the testing process, the new output data with known target is compared to the predicted data with the known targets. We here are using the credit risks as low, medium and high as the target values which are defined from the start and we classify and compare these with the individual credit risk to determine the viable candidate.

Here we cannot use the clustering or regression approach as the we don’t have all the data in numerical or continues for due to which the regression is not possible and coming to the clustering method, it involves collection of similar data but here we have different data of a single user and we need to classify it to predict the credibility of the user. Clustering might be useful if we are to do the customer profiling instead of deciding the loan applicants.

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.