Data Mining Assignment 1

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

**Problem Statement:** Given previous month’s data of various employees, check the upcoming employee performance

**1. What the data is.**

Here, data is the performance of employees from past 1 year like his daily target achieved, attendance, quality score, error score, past awards, leaderboard score, exam score. Each data is required to check if a particular employee will perform well or will not perform good in upcoming months.

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

Here we can use data mining to get necessary features, if any data is missing, using different techniques to fill up the data, and various data mining techniques to predict values of upcoming months, so that if any employee is underperforming, we can take necessary precautions/tasks.

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

We can use regression model to predict data. Since we are given previous months data, we can use that data to train and test the data. We use regression model to get a hypothesis and use that hypothesis to test our test set. Using different techniques or using different features we can get MSE and select the one with lease MSE.

If our model is good, then we can predict if a employee is underperforming, that will help the company to take necessary actions.

**4. Name one type of data mining that you think would not be relevant, and describe briefly why not.**

Here, classification model is not relevant, because, since we are given linear data from past several months, and linear data is not suitable for classification model. Here we are predicting employee score so that we can come to a conclusion that if the score is not upto the mark, he is underperforming.

**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.**

In the above example, clustering is not relevant because the given data set is in linear fashion, and clustering the data will do us no good. It will not predict correctly. Clustering can be done if there is data which needs to be separated so that similar data can be clustered together. Which will help us to determine, to which the given test data can be divided.