A. Describe the dataset

This dataset is related to the record of the employees in the company. This dataset contains 14999 rows and 10 columns. The dataset describes the satisfaction level of the company, last evaluation of their performance, number of projects they done, their average monthly working hours, the number of years spent in the company, whether the employee had a workplace accident, whether the employee left the workspace, their last promotion, the sales and the salary level. We had found out that some of the column names are too ambiguous, for example, the column "last_evaluation" does not give a clear meaning. Besides, the data unit of the column is ambiguous, for example, the column name "time_spend_company" is recorded in integers values (3,4,5), however, it could represent anything, and in this case, represents unit years. Another problem that we found in this dataset is that true and false are represented as 1 and 0 instead.

B. Insight

The insight that we wish to extract from this dataset is the next employee who will likely leave the company. We want to find out how to keep employees satisfied.

C. Data Mining Technique

The data mining technique that will be relevant is classification, by analysing the dataset we can find out what kind of employees are likely to leave the company. For example, we can look in the satisfaction level, the number of projects and salary rate to determine whether the employee will leave or stay in the company. Another data mining technique that we wish to apply is association rule. For example, we can discover someone with the different level of salary across the different department that will have a high likelihood of leaving the company.

D. Data quality issues

After checking the dataset against the six data qualities dimension which is uniqueness, timeliness, completeness, consistency, accuracy, and validity, we found out that the dataset contains duplicate entry and may need processing. Besides, some of the column names do not properly describe its content such that:

- time_spend_company - the column name didn't specify the data unit
- last_evalution – we do not know what the column name is trying to imply.
- Sales – the column name is not relevant to the values.
- average_montly_hours – Spelling Error