DM Practical 1

`ROLL No.`: 21BCE020
`AIM`: Data Domain selection and Identification of Characteristics of selected Dataset of different Formats

---

Selection of data domain

→ Clothing Industry

Definition of data domain

- Supply chain management: Management of systems and improving productivity by minimizing excess inventory, optimizing flow of materials from suppliers to manufacturers and improving schedule management leads to reduced delays and increased productivity.
- Collaboration: Effective collaboration and communication at different levels of management influence the design operations and reduces delays.
- Incentive-driven development: Training programs which provide incentives to employees can help workers be updated and motivated which contributes significantly to productivity.

Data source

- Productivity Prediction of Garment Employees

Objective

- To predict growth of a firm on the basis of productivity displayed by employees.
- Identify the behavioural patterns of employees through incentives
- To predict how productivity affects market sales

Define the selection of fields

→ Features selected for further processing are targeted_productivity, actual_productivity, over_time, incentive.

Characteristic and behaviors (distribution and inference) of data for each selected field

- Targeted_productivity: Shows the percentage of productive work to be done by individual employees
- Actual_productivity: Shows the percentage of work actually performed by employees
- Over_time: Amount of overtime of each team in minutes

- Incentive: Represents the incentives provided which motivates employees

---

KDD Processes

- Data cleaning: The dataset requires thorough examination for missing values and outliers.
- Data integration: The dataset under observation does not require integration.
- Data selection: targeted_productivity, actual_productivity and incentive can be used to predict productivity of employees and evaluate model accuracy.
- Data transformation: The dataset can using scaling and normalization of values.
- Data mining: From the dataset, we can get information about productivity of employees and the idle time of employees. Classification models can be used on the dataset.
- Pattern evaluation: We can observe the productivity of employees and compare it to the target productivity. Furthermore, we can observe idle time of employees and the number of employees who are idle due to several reasons.
- Knowledge representation: Productivity and idle time of employees can be represented in line graphs, scatterplots and heatmaps.