CRISP-DM stage three – data preparation

Select data

Task

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

Output

- **Rationale for inclusion/exclusion** – list the data to be included/excluded and the reasons for these decisions.

Clean data

Task

Raise the data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.

Output

- **Data cleaning report** – describe what decisions and actions were taken to address the data quality problems reported during the Verify Data Quality task of the Data Understanding phase. Transformations of the data for cleaning purposes and the possible impact on the analysis results should be considered.

Construct data

Task

This task includes constructive data preparation operations such as the production of derived attributes or entire new records or transformed values for existing attributes.

Outputs

- **Derived attributes** – these are new attributes that are constructed from one or more existing attributes in the same record. Example: area = length * width.
- **Generated records** – describe the creation of completely new records. Example: Create records for customers who made no purchase during the past year. There

was no reason to have such records in the raw data, but for modelling purposes it might make sense to explicitly represent the fact that certain customers made zero purchases.

Integrate data

Task

These are methods whereby information is combined from multiple databases, tables or records to create new records or values.

Outputs

- **Merged data** – merging tables refers to joining together two or more tables that have different information about the same objects. Example: a retail chain has one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarized sales data (e.g., profit, percent change in sales from previous year), and another with information about the demographics of the surrounding area. Each of these tables contains one record for each store. These tables can be merged together into a new table with one record for each store, combining fields from the source tables.
- **Aggregations** – aggregations refer to operations in which new values are computed by summarizing information from multiple records and/or tables. For example, converting a table of customer purchases where there is one record for each purchase into a new table where there is one record for each customer, with fields such as number of purchases, average purchase amount, percent of orders charged to credit card, percent of items under promotion etc.