CRISP-DM

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding

The first stage of the framework is to develop a business understanding. It has two steps:

- l Determine the business objective
- l Identify the goal of the data analysis

2. Data understanding

This stage comprises four key steps to understand the available data, and identify new relevant data in order to solve the business problem.

- Collect relevant data
- Describe data – for explicit information
- Explore data – for implicit insights
- Verify data quality – to remove errors

3. Data preparation

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks:

- Select data: Determine which data sets will be used.
- Clean data: A common practice is to correct, impute, or remove erroneous values.
- Construct data: Derive new attributes that will be helpful.
- Integrate data: Create new data sets by combining data from multiple sources.
- Format data: Re-format data as necessary.

4. Modeling

- The first task is to understand the problem domain and select the appropriate family of models that is suitable for solving the problem at hand.
- The second task is to select appropriate algorithms for creating the model from the chosen family of models.

5. Evaluation

- The predictive models can be tested to assess their effectiveness in solving the problem. This is the fifth stage of the framework – model evaluation.
- Modelling and evaluation together is an iterative process in which the models are tweaked until satisfactory evaluation results are obtained.

6. Deployment

This is the last stage of the framework, where the model is translated into a business strategy. Business data is fed into the model and the model results are used to inform business decisions on an on-going basis.