# Data Science Methodology (Crisp-DM)

## Contents

# 1 What is CRISP-DM

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral. CRISP-DM is comprised of six steps with an entity that has to implement in order to have a reasonable chance of success.

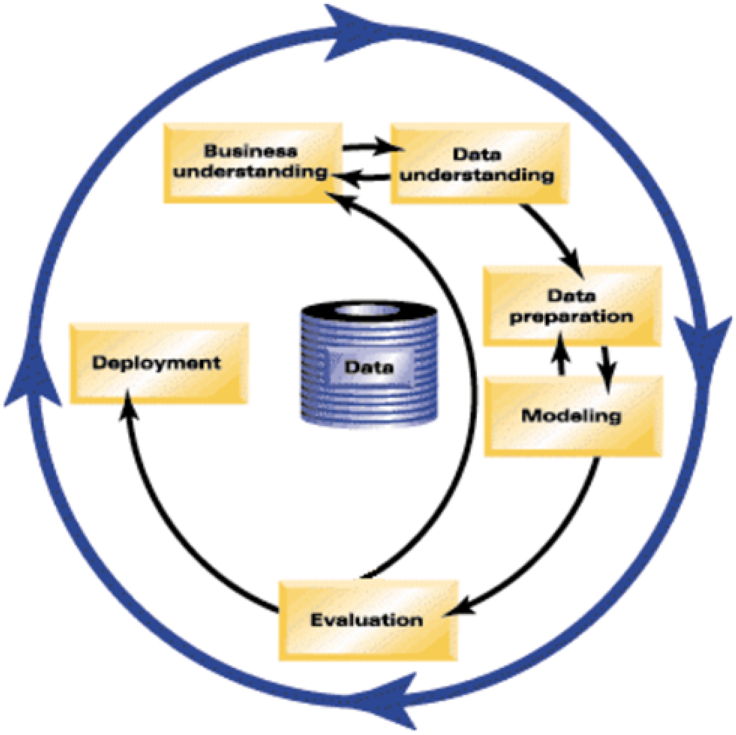

Figure 1: Crisp-DM

1. **Business Understanding**

   This stage is the most important because this is where the intention of the project is outlined. Foundational Methodology and CRISP-DM are aligned here. It requires communication and clarity. The difficulty here is that stakeholders have different objectives, biases, and modalities of relating information. They don't all see the same things or in the same manner. Without clear, concise, and complete perspective of what the project goals are resources will be needlessly expended.

2. **Data Understanding**

   Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

3. **Data Preparation**

   Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.

4. **Modeling**

   Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Model selection is an art and science. Both Foundational Methodology and CRISP-DM are required for the subsequent stage.

5. **Evaluation**

   The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

6. **Deployment**

   In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model.These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

CRISP-DM is a highly flexible and cyclical model. Flexibility is required at each step along with communication to keep the project on track. At any of the six stages, it may be necessary to revisit an earlier stage and make changes. The key point of this process is that it's cyclical; therefore, even at the finish you are having another business understanding encounter to discuss the viability after deployment. The journey continues.

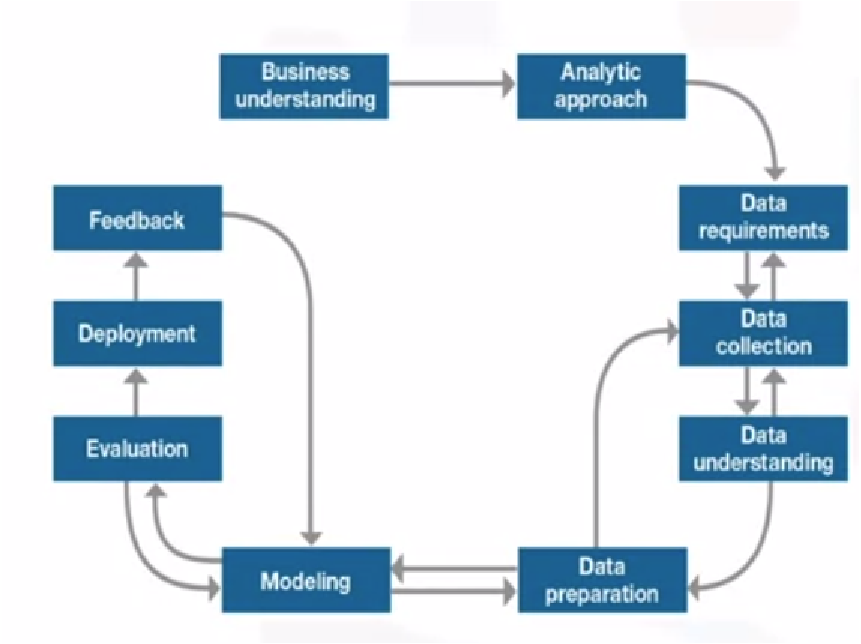

Figure 2: Advance Crisp-DM

## 2 Summary

- The need to understand and prioritize the business goal.
- The way stakeholder support influences a project.
- The importance of selecting the right model.
- When to use a predictive, descriptive, or classification model.
- The significance of defining the data requirements for your model.
- Why the content, format, and representation of your data matter.
- The importance of identifying the correct sources of data for your project.
- How to handle unavailable and redundant data.
- To anticipate the needs of future stages in the process.
- The importance of descriptive statistics.
- How to manage missing, invalid, or misleading data.
- The need to clean data and sometimes transform it.
- The consequences of bad data for the model.
- Data understanding is iterative; you learn more about your data the more you study it.
- The difference between descriptive and predictive models.
- The role of training sets and test sets.
- The importance of asking if the question has been answered.
- Why diagnostic measures tools are needed.
- The purpose of statistical significance tests.
- That modeling and evaluation are iterative processes.
- The importance of stakeholder input.
- To consider the scale of deployment.
- The importance of incorporating feedback to refine the model.
- The refined model must be redeployed.
- This process should be repeated as often as necessary.