Data Mining and Text Mining

Rei Sanchez-Arias, Ph.D.
rsanchezarias@floridapoly.edu

Cross-Industry Standard Process for Data Mining (CRISP-DM)

The **CR**oss **I**ndustry **S**tandard **P**roc ess for **D**ata **M**ining (*CRISP-DM*) is an open and structured process model. Founded in the late 90s to standardize data mining processes across industries, it has since become a common methodology for data mining, analytics, and data science projects.

Learn more here:
https://www.datascience-pm.com/

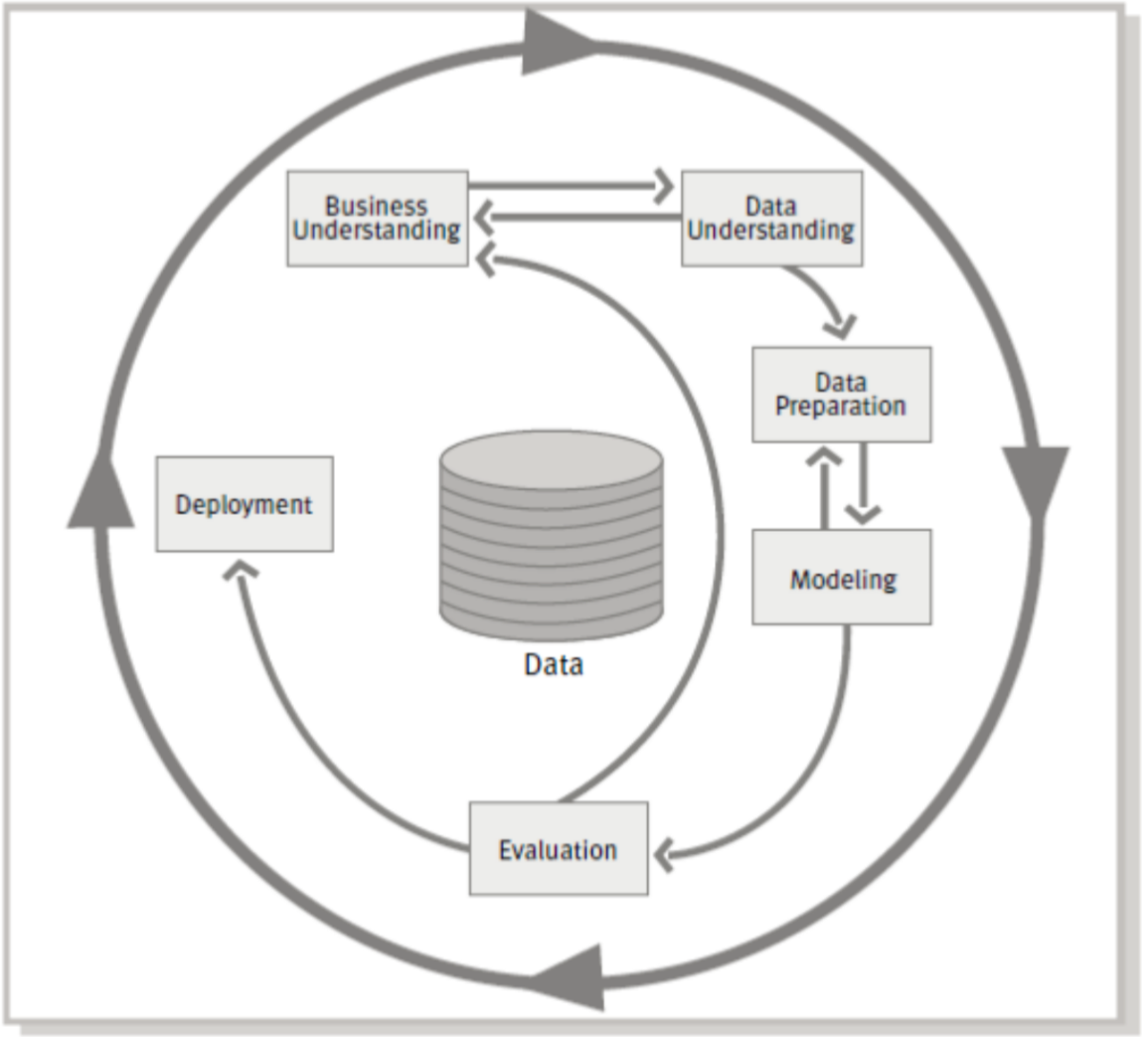

Image credit: CRISP-DM 1.0 guide [crisp-dm.org website]

Data Understanding

- Collecting the data.
- Describing the data.
- Exploring the data.
- Verifying the data quality.

This step is the classic case of Extract, Transform, Load (**ETL**)

Data Preparation

- Selecting the data.
- Cleaning the data.
- Constructing the data.
- Integrating the data.
- Formatting the data.

Modeling, Evaluation and Deployment

Modeling

- Selecting a modeling technique.
- Generating a test design.
- Building a model.
- Assessing a model.

Evaluation

- Evaluating the results.
- Reviewing the process.
- Determining the next step

Deployment

- Deploying the plan.
- Monitoring and maintaining the plan.
- Producing the final report.
- Reviewing the project.

About Data Preparation

Bringing features onto the same scale

Feature scaling is a crucial step in our preprocessing pipeline that can easily be forgotten.

There are two common approaches to bringing different features onto the same scale: **normalization** and **standardization**. (Those terms are often used quite loosely in different fields, and the meaning must – sometimes – be derived from the context)

Normalization

Most often, normalization refers to the **rescaling** of the features to a range of [0, 1], which is a special case of *min-max* scaling. To normalize our data frame, we can simply apply the min-max scaling to each feature column, where the new value $x_{norm}^{(i)}$ of sample $x^{(i)}$ can be calculated as:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

x_{min} is the smallest value in a feature column, and x_{max} is the largest value.

Original Data

Sepal.Length		Sepal.Width	
Min.	:4.300	Min.	:2.000
1st Qu.	:5.100	1st Qu.	:2.800
Median	:5.800	Median	:3.000
Mean	:5.843	Mean	:3.057
3rd Qu.	:6.400	3rd Qu.	:3.300
Max.	:7.900	Max.	:4.400

Normalized Data

Sepal.Length		Sepal.Width	
Min.	:0.0000	Min.	:0.0000
1st Qu.	:0.2222	1st Qu.	:0.3333
Median	:0.4167	Median	:0.4167
Mean	:0.4287	Mean	:0.4406
3rd Qu.	:0.5833	3rd Qu.	:0.5417
Max.	:1.0000	Max.	:1.0000

Standardization

Using standardization, we **center** the features columns at mean 0 and standard deviation 1. Notice, that standardization maintains useful information about **outliers** and makes algorithms less sensitive to them in contrast to mix-max scaling, which scales the data to a limited range of values. The procedure of standardizations follows:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

μ_x is the sample mean of a particular feature column, and σ_x the standard deviation.

Original Data

Sepal.Length	Sepal.Width
Min. :4.300	Min. :2.000
1st Qu.:5.100	1st Qu.:2.800
Median :5.800	Median :3.000
Mean :5.843	Mean :3.057
3rd Qu.:6.400	3rd Qu.:3.300
Max. :7.900	Max. :4.400

Standardized Data

Sepal.Length	Sepal.Width
Min. :-1.86378	Min. :-2.4258
1st Qu.:-0.89767	1st Qu.:-0.5904
Median :-0.05233	Median :-0.1315
Mean : 0.00000	Mean : 0.0000
3rd Qu.: 0.67225	3rd Qu.: 0.5567
Max. : 2.48370	Max. : 3.0805