Table of Contents

Introduction

Why CRISP-DM?

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide data mining efforts in a Knowledge Discovery Process. Whilst other tech-driven methodologies are available, the CRISP-DM methodology has gained popularity over the years due to its distinctive business-oriented approach[1]. As opposed to other discovery methods, such as SEMMA (Sample, Explore, Modify, Model and Assess) or KDD (Knowledge Discovery in Databases), CRISP-DM is characterized by a continuous and iterative rapport between data mining and business acumen and their respective responsible persons[2]. With this all-round integration of both data and business into the equation, CRISP-DM ensures that the Knowledge Discovery Process delivers meaningful business value.

- As a methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.
- As a process model, CRISP-DM provides an overview of the data-mining life cycle.

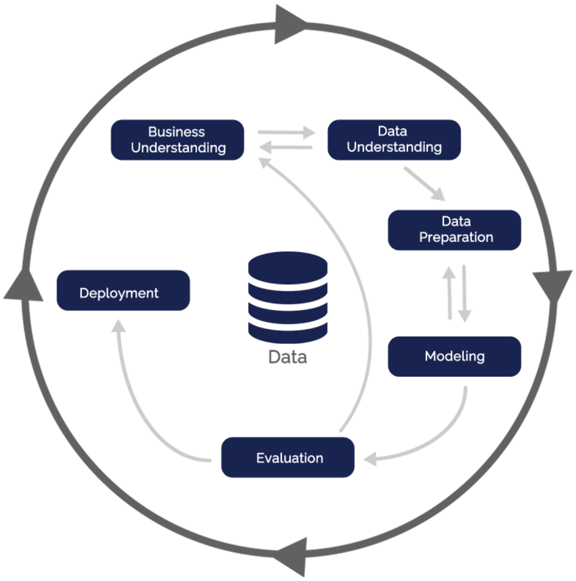


Figure 1: CRISP-DM Methodology Life Cycle Schematic

The initial phase aims to obtain a clear understanding of the desired objectives from a business perspective. Consequently, the business goal needs to be converted to a data problem and result in a project plan. Through this process, a clear picture of the business goals is drawn so that the analytics plan can be customized and correctly tailored.

[1] https://analyticsindiamag.com/crisp-dm-data-science-project/
[2] Azevedo and Santos (2008)

The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict, but customizable and flexible to fit the purpose of the given task at hand, and its particular needs. In fact, most projects move back and forth between phases as necessary[3].

What can you expect from this handbook?

This document provides a step-by-step guideline to successfully complete a Knowledge Discovery Process through CRISP-DM and deliver meaningful business value. The guideline also highlights key areas of decision-making and the factors which need to be considered in order to guarantee the best outcome possible. Furthermore, the handbook is accompanied by a business scenario to illustrate the application of the theory and the procedures. Specifically, the problem that this handbook will focus on is the issue of catastrophic pipeline failures.

Damaged pipes impede supplying water to customers. The disruption of this service causes significant financial losses for the asset holding company. On the one hand, the interruption in the provision of water results in losing customers. On the other hand, as the company loses one of its largest assets due to a rupture, the company's value decreases. In addition, ruptured pipes imply a significant waste of water that is not only detrimental to the environment but also to the people living in the region.

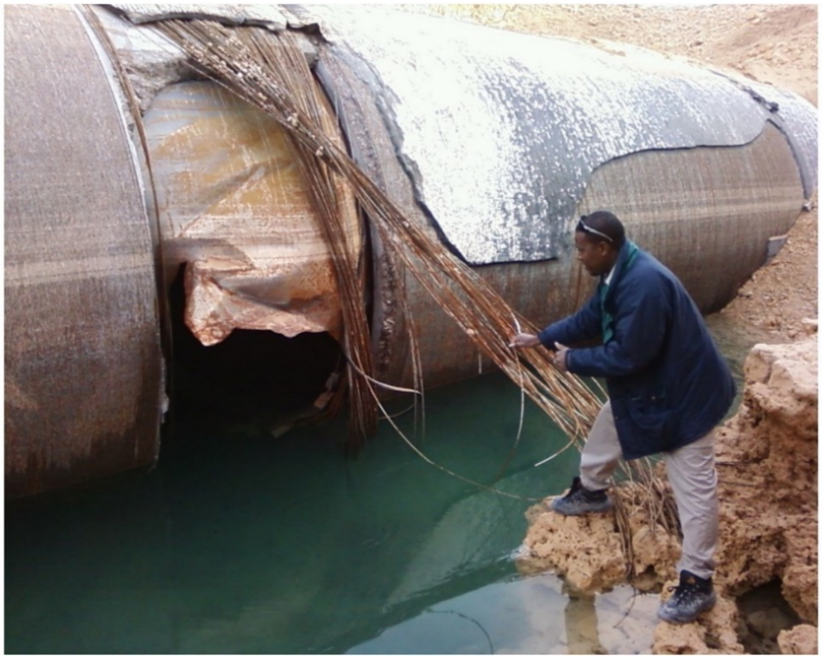


Figure 2: Pipeline Rupture on the North African Coast disrupting service to entire cities

[3] IBM SPSS Modeler CRISP-DM Guide

This purpose of the example is to act as a vehicle throughout all the phases and facilitate a better understanding and assimilation of the concepts.

1. Business Understanding

1.1 Business Understanding Overview

The initial phase aims to obtain a clear understanding of the desired objectives from a business perspective. Consequently, the business goal needs to be converted to a data analytics goal problem definition[4].

It is critical to understand what an organization expects to gain from data analytics project and to involve as many key people as possible prior to commencing any effort. Getting to know the business reasons for the data analytics effort helps to ensure that everyone is on the same page before expending valuable resources[5].

1.2 Determining Business Objectives

Today, new advancements in technologies and data analytics are helping utilities build asset management programs using a risk-based approach to pipeline condition assessment with the lowest financial impact.

There is no one-size-fits-all approach to assessing pipelines. An approach should be tailored within the context of the organization's risk tolerance while taking into consideration the material, diameter, and past failure history. The approach can range from do-nothing to a full in-line inspection making targeted repairs and be progressive in nature. Many different methods and technologies can be combined, to provide data and information to make decisions and prioritize maintenance repairs and shutdowns[6].

Utilities often used indirect methods of assuming the condition of the pipeline or replaced based on age and consequence of failure, not on the actual condition of the infrastructure. By applying the CRISP-DM methodology, a risk-based approach can be developed, to best fit an organization's goals and risk tolerance, optimize capital expenditures, prevent failures and increase confidence and level of service.

1.2.1 Compiling Business Background

Prior to commencing the project, it is necessary to understand the business setting within the context and understand what the project is aiming to achieve and what problems need to be solved.

[4] Crisp-DM Methodology Luchtvarfeiten.nl
[5] IBM SPSS Modeler Crisp-DM Guide
[6] https://puretechltd.com/water-and-wastewater/

After reviewing and understanding the project scope the following tasks need to be completed:

- Map organizational and project structure identifying internal and external stakeholders, as well as the department(s) impacted by the project.
- Identify key individuals, their roles and their responsibilities.
- Define the problem that needs to be solved and clarify prerequisites.
- Describe possible available solution as part of the project roadmap.

1.2.2 Defining Business Objectives

Once a general understanding has been achieved, it will be necessary to commence discussions with key business stakeholders or sponsors initiating the project, in order to define the specific objectives that need to be achieved.

After the initial kick-off meeting, realistic, clear and attainable goals need to be set. Prior to proceeding with the project plan, the following tasks need to be completed:

- Describe the problem to be solved, specifying precise business questions.
- Determine business requirements and specify expected gains and benefits.

1.2.3 Business Success Criteria

Once business goals have been clarified, it is important key to determine what the success criteria for the projects are. It will be necessary to define what will make the project a success and how it will be measurable in terms of a business objective.

There are two main criteria to consider:

- Technical success: Succeeding in completing the project scope from a technical standpoint; understanding the problem and being able to make data-driven decisions.
- Business success: Achieving technical success with minimal risks and within a specific budget and timeline.

1.3 Assessing the Situation

Once the business goals and success criteria have been defined, it is necessary to perform an assessment of the current situation in order to determine how to proceed with the project plan. It is critical to consider the following:

- What data are currently available?
- What personnel is indispensable or the project?
- What are the projects risks?
- Can a contingency plan be formulated to address the project risks?

1.3.1 Resource Inventory

Once the situation has been assessed, it will be required to compile a list of available inventory resources and consider the following:

- What tools and equipment will be indispensable?
- What staff members are available, including subject matter experts and administrators?

1.3.2 Requirements, Assumptions and Constraints

As part of the assessment process, it is important to realistically identify liabilities and consider the following:

- Are tools and equipment available for use during the project timeline?
- Are staff members aligned with the project schedule?
- Are there any legal or contractual restrictions regarding data collection?
- Are there financial constraints that will affect time and effort dedicated to the project budget?
- Are there assumptions on data quality?
- What are the expectations of internal and external stakeholders

1.3.3 Risks and Contingencies

It is equally important to identify project risks in coordination with the project team and stakeholders, document them and develop mitigation strategies and contingency plans for each one.

Consider the following:

- Scheduling risks – what if the project requires a longer timeline than originally planned for.
- Financial risks – what if financial problems appear during the project lifecycle and funding becomes scarce.
- Data risks – what if there are errors in data, the data is of poor quality or false positives exist in the end results.

1.3.4 Terminology

In order to facilitate good communication and clarity, the compilation of a glossary with technical terms is recommended. A list of terms can be compiled in coordination with the project team and based on working experience of past projects. The glossary can be attached as an appendix to the final report for reference.

1.3.5 Cost Benefit Analysis

As part of the assessment process, it is critical to consider budget, revenue, projected margin and essentially deduce a bottom that will be the driving factor throughout the project lifecycle and inform project activities in terms of time and materials expended.

A preliminary budget should consider:

- Planning and mobilization costs
- Operating and data collection costs
- Reporting costs

1.4 Determining Data Mining Goals

Once business objectives have been defined and clarified, they should then be translated into a data problem. Reducing capital maintenance costs for example can be translated into identifying and locating problem areas on the pipeline and being able to make data driven decisions and perform proactive targeted maintenance programs and repairs that address problem areas alone instead of waiting for a catastrophic failure or blindly performing maintenance on the entire asset. It is critical that business problems are meaningfully translated into data problems and resolved hand in hand.

1.4.1 Data Mining Goals

Input from the operations and data analysis teams, as well as subject matter technical experts is taken into consideration in order to come up with a sound technical solution. The data mining goals should include:

- A specific technical solution along with roadmap.
- A specific timeline and schedule to deliver these goals.
- Projected project outcomes and specific type of results.

1.4.2 Data Mining Success Criteria

The criteria that will determine the project's success must also be defined in technical and specific terms. It is important to consider the following:

- The methodology for data collection and analysis.
- Introducing benchmark criteria to evaluate results.
- The method of evaluation of the project outcomes.

1.5 Producing a Project Plan

The project plan is the master reference document that documents, defines and informs all project activities, as well as project goals, resources, operational tasks, project risks, risk mitigation strategies and contingencies, roles and responsibilities, project schedule and a timeline for deliverables.

The project plan should consider input from all team members involved, and at the same time take stakeholder concerns and restrictions into consideration. A kick-off meeting should be held prior to its formulation in order to collectively review the effort and necessary requirements.

The final project plan should include a clear flow of activities, time estimates for each activity, a list of resources for each activity, staff members responsible for carrying out each activity and a clear timeline for internal and external deliverables.

2. Data Understanding

2.1 Data Understanding Overview

The data understanding phase aims to increase familiarity with the available data. The initial data is collected from the project resources and further examined to identify the main characteristics of the data. The data requires further exploration to address the specific questions. Lastly, the data needs to be assessed on its quality[7]. A project may contain multiple different datasets. Each of these datasets are examined to identify exactly what they contain and how it ultimately impacts the project.

2.2 Collecting Initial Data

There are two (2) categories of data that will be used to compile the pipeline condition assessment: existing data and additional inspection data.
Existing data may come from a variety of sources including[8]:

Pipeline Plan and Profile Drawings: Engineering drawings prepared in CAD that combine a plan view of the asset pipeline as well as a profile XY view of the elevation.

Pipeline Lay Sheets: Blueprint drawings and tables showing the exact route of the pipeline and the specifications of each individual pipe stick.

Pipe Design and Manufacturing Specifications: A table prepared during the design phase of any project. It provides the appropriate selection, specification and material grade of pipe and piping components

Operational Pressure Data: A table with historical logs of actual pipeline pressure recorded at different intervals during a given time period.

History of Past Failures and Repairs: A log or records that indicate and describe past failures, their location and type of repair.

Additional inspection data may include:

GIS Mapping Data: High accuracy geographical information data that captures the location of known features.

Visual and Sounding Inspection Data: Data logged by the field crew during internal pipe inspections. Visual data would include observations such as cracks in the surface or degradation at the joints. Sounding data would include observations derived from tapping the pipe surface to identify hollows and possible delamination.

[7] Crisp-DM Methodology Luchtvarfeiten.nl
[8] Sagiannos (2017)

Acoustic Leak Detection Data: Acoustic inline inspection data identifying the location of uncontrolled leaks and gas pockets.

Electromagnetic Structural Inspection Data: Electromagnetic inline inspection data identifying areas of corrosion on the pipe wall and their location.

Transient Pressure Monitoring Data: Actual pressure monitoring data reporting at predefined intervals at normal operating conditions. When a transient is detected, the sample rate is increased in order to provide accurate readings to plot the entire transient event.

Manufacturing Date: Indicates the specific material used and the standard followed.

2.2.1 Data Collection Report

Once the data has been gathered, a data collection report should be written and distributed among the project team and stakeholders. The data collection report should include:

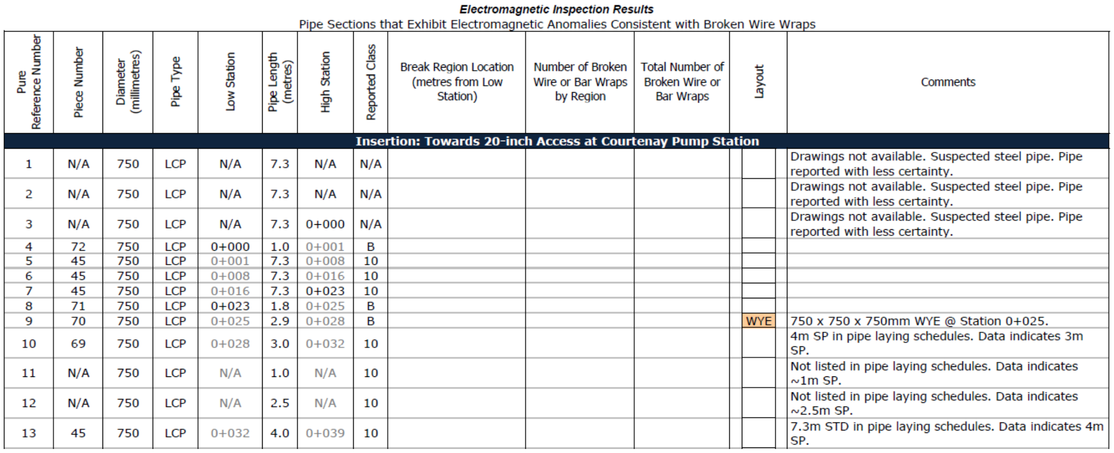
- Data requirements: the type data required to complete the project, and whether it is available or not.
- Selection criteria: which attributes are necessary, which attributes are irrelevant and which attributes can be combined together to provide insights. Additionally, the selection criteria includes identifying which files/tables are of interest, what data within the file/table are of interest and what background history is relevant

It is possible to combine this report with material that needs to be reported in the next steps of this phase in order to produce a consolidated document that will guide work through the data preparation phase.

2.3 Describing Data

The data sets acquired should be compiled, and ultimately combined to create a single table that represents the asset pipeline, pipe stick by pipe stick. Each row of the table would represent one pipe stick, forming a sequence that describes the pipeline, while columns of the table would represent attributes of the given pipe stick.

Attributes of each pipe will include an arbitrarily assigned sequential reference number, station numbers as derived from Lay Sheets and Plan & Profile Drawings; pipe type, length and diameter as derived from Pipe Manufacturing Specifications; pressure as derived Operational Pressure Data; notes indicating past failures and repairs (if any), as derived from Historical Records. Additionally, the table should include available columns for attributes such as quantified amount of distress, as derived from Electromagnetic Inspection Data; presence of leaks and their exact location (barrel or joint); GPS coordinates on known features as derived from GIS mapping data; notes on visual and sounding observations, as derived from Visual and Sounding Inspection data.



Electromagnetic Inspection Results

Pipe Sections that Exhibit Electromagnetic Anomalies Consistent with Broken Wire Wraps

Pure Reference Number	Piece Number	Diameter (millimetres)	Pipe Type	Low Station	Pipe Length (metres)	High Station	Reported Class	Break Region Location (metres from Low Station)	Number of Broken Wire or Bar Wraps by Region	Total Number of Broken Wire or Bar Wraps	Layout	Comments
Insertion: Towards 20-inch Access at Courtenay Pump Station												
1	N/A	750	LCP	N/A	7.3	N/A	N/A					Drawings not available. Suspected steel pipe. Pipe reported with less certainty.
2	N/A	750	LCP	N/A	7.3	N/A	N/A					Drawings not available. Suspected steel pipe. Pipe reported with less certainty.
3	N/A	750	LCP	N/A	7.3	0+000	N/A					Drawings not available. Suspected steel pipe. Pipe reported with less certainty.
4	72	750	LCP	0+000	1.0	0+001	B					
5	45	750	LCP	0+001	7.3	0+008	10					
6	45	750	LCP	0+008	7.3	0+016	10					
7	45	750	LCP	0+016	7.3	0+023	10					
8	71	750	LCP	0+023	1.8	0+025	B					
9	70	750	LCP	0+025	2.9	0+028	B				WYE	750 x 750 x 750mm WYE @ Station 0+025.
10	69	750	LCP	0+028	3.0	0+032	10					4m SP in pipe laying schedules. Data indicates 3m SP.
11	N/A	750	LCP	N/A	1.0	N/A	10					Not listed in pipe laying schedules. Data indicates ~1m SP.
12	N/A	750	LCP	N/A	2.5	N/A	10					Not listed in pipe laying schedules. Data indicates ~2.5m SP.
13	45	750	LCP	0+032	4.0	0+039	10					7.3m STD in pipe laying schedules. Data indicates 4m SP.

Figure 3: Example of consolidated data table with rows as pipe sections and attributes as columns[9]

2.3.1 Data Description Report

A data description reports needs to be produced summarizing how the collected data has been put together and answer questions such as:

- What data types are present?
- Have sources for each dataset been identified?
- What is the format of each collected dataset?
- What tool/method was used to capture the data or where was the data derived from?
- Is the data sufficient to achieve the target business goal?
- Have basic statistics for key attributes been computed and what insights towards business goals can be deduced?
- Can attributes relevant to business goals and project scope be prioritized? If not, is it possible to extrapolate data in order to reach a value adding result?

2.4 Exploring Data

Once the datasets have been compiled and consolidated, conducting an in initial exploration is useful in order to look for errors in the data. Such analyses can help formulate hypotheses and shape the data transformation tasks that will take place under the data preparation.

Examples would include reviewing overlapping datasets, such as comparing pipe lay schedule to plan and profile drawings, electromagnetic and acoustic data in order to identify discrepancies and anomalies. These need to be discussed project stakeholder to further understand the reason of their existence.

[9] Courtenay Pump Station Condition Assessment Report, Comox Valley Regional District, Sagiannos 2017

2.4.1 Data Exploration Report

As tables and graphs are created in this phase of the methodology hypotheses start to form about which business goals the data can answer. It is therefore necessary to compile a Data Exploration Report that addresses the following questions:

- What assumptions have been formed about each dataset?
- Which attributes are more relevant for further analysis in order to achieve the given scope?
- Are there any discrepancies between overlapping datasets such as lay sheets and electromagnetic inspection data? If yes, how can these discrepancies be reasonably explained?
- Have the explorations revealed new information (or lack of) about the project? Have the initial hypotheses changed based on these findings or have the business goals been affected? Can the project scope still be achieved?
- Can any of the explored datasets be used to serve future assessments of the subject pipeline asset? Are there data for sections of the pipeline beyond the given contract scope that can be stored and used for future projects?

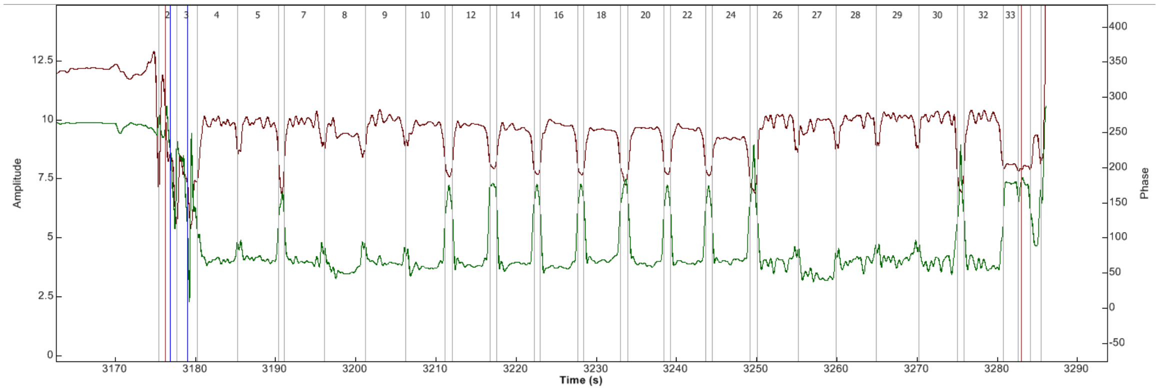


Figure 4: Electromagnetic data sample - each pipe section can be clearly identified due to phase shifting in response and can be overlaid with lay sheet information to identify exact locations of distress[10]

2.5 Verifying Data Quality

Data quality verification can typically be conducted during the data description or exploration phases in parallel with other tasks. As datasets are collected from different sources they are far from perfect.

Possible issues could include:

[10] Fish Hatchery Pump Station Condition Assessment, Halifax Water, Sagiannos 2014

Measurement errors: It is possible that errors have been introduced into different datasets from poor or inconsistent recording of parameters during the installation of the assets or in field measurements due to wrong tool setup or not following standard data collection procedures.

Missing data: Missing data could be non-existent lay sheets, missing plan and profile drawings or pipeline features or appurtenances that have been added after installation and not been recorded.

Data errors: Additional datasets collected in the field may contain errors, noise or masked data. These errors could result to attributes having inconsistent or unusual parameters or even false positives.

2.5.1 Data Quality Report

Based on observations recorded during the data quality verification phase, a data quality report should be compiled. The report should aim to summarize data quality concerns and consider the following:

- Have missing attributes been identified? If yes, is there a meaningful explanation for the missing information?
- Are there any inconsistencies that can cause problems in later stages of the process?
- Have anomalies or noise events been sufficiently explored to determine what they are?
- Have attribute values been reviewed to ensure they make sense?
- Can bad or noisy data be analyzed?
- Can data quality issues be addressed?

3. Data Preparation

3.1 Data Preparation Overview

The data preparation phase involves all activities carried out to construct the final set of data from the raw initial data of the previous phase. Once a solid understanding of what data does or does not exist has been developed, the data is prepared and analyzed in a way to make it useful. The data needs to be cleaned in order to correct for any inaccurate records. The set of data might still require a few adjustments, such as derivation of attributes by certain calculations and the generation of completely new records. On top of that, the data is integrated whereby information of multiple tables is combined. Ultimately, the data might require any formatting transformations, in order to properly feed the data to the modelling tools[11].

3.2 Selecting Data

Based upon the initial data collection conducted, data relevant to the project scope and business goals can be selected. The decision about which data to select are typically made during the business understanding phase of the project, whereupon the project scope is determined. The quality of each particular dataset should also be considered in order to determine the validity of the project outcomes and manage internal and external stakeholder expectations accordingly.

There are two (2) ways to select data:

- **Select rows:** Based upon the project scope we can determine pipe length and therefore station numbers which correspond to specific pipe sticks. As each pipe stick represents a row in our consolidated data table, a decision determined by the contractual scope of work needs to be made in order to include all available data or specific sections. In case that sections of the pipeline for which data is available is beyond project scope, then the table rows that correspond to the given scope should be excluded.
- **Select attributes:** Based upon the project contractual scope we can determine which attributes are relevant to process in order to achieve the desired outcome. The dataset that inform the table attributes should relevant to the pre-determined project outcomes and selected accordingly. If for example pressure data are beyond the project scope then the pressure dataset does not need to be included in further analysis and reporting.

[11] Crisp-DM Methodology, Luchtvarfeiten.nl

3.3 Cleaning Data

The data cleaning phase of the project involves reviewing the datasets that have been selected for analysis at a closer level. The cleaning process is applied in an effort to address and correct issues noted in the data quality report. Similar to the data quality verification phase, these issues would include:

Measurement errors: It is possible that errors have been introduced into different datasets from poor or inconsistent recording of or field measurements. For problematic items, it is important to review all possibly overlapping datasets and attempt to align them as much as possible to the extent that it makes actual sense.

Missing data: Missing data could be compensated by extrapolating values and/or compensating with data from overlapping datasets. Where such an approach is not possible, assumptions based on existing information and past projects should be made.

Data errors: Datasets containing errors or noise can be corrected here. This task could involve applying filters for background noise or manipulating data in order to make their content clearer.

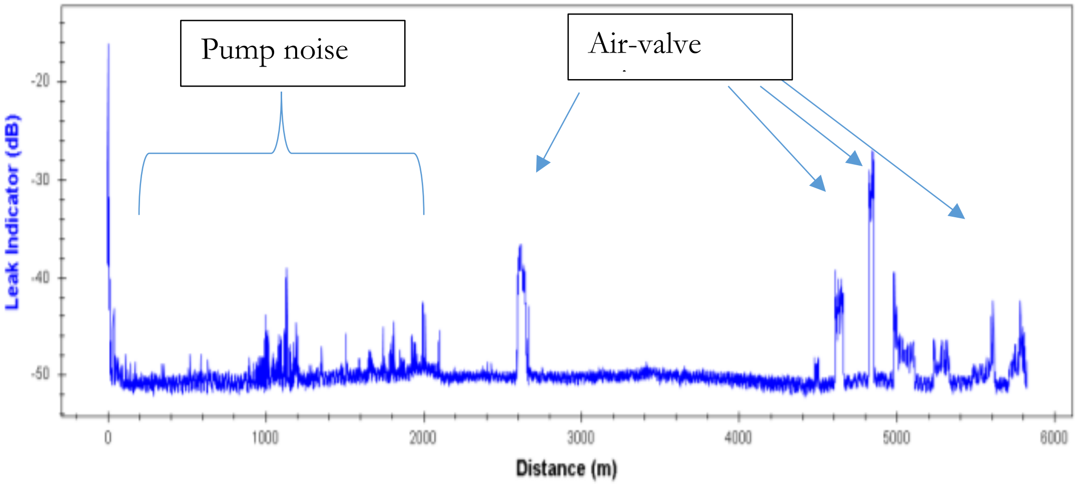


Figure 5: High Level Overview of Acoustic Leak Detection Data containing distortion from pump noise and spikes from air-valve noise[12]

3.3.1 Data Cleaning Report

Data cleaning efforts should be logged in a data cleaning report in order to track alterations to subject datasets, assumptions and hypotheses.

[12] Courtenay Pump Station Condition Assessment Report, Comox Valley Regional District, Sagiannos 2017

Examples of efforts included in the report are:

- What discrepancies existed in the data and how were they handled.
- What types of noise were present in the data and how were they removed.
- Is it possible to salvage noisy data and to what extent.
- What techniques or assumptions were used to make up for missing data.

3.4 Constructing New Data

During the project lifecycle, it may be necessary to construct new data. A good reason for this could be a change order that augments the project scope. For this purpose, new data can be constructed in two (2) ways:

- Adding new rows: increasing the asset pipeline distance by adding pipe sticks
- Deriving attributes (adding columns): this could involve adding datasets to the analysis of the asset pipeline such as leak detection or pressure data that were initially beyond project scope and not selected. Additional attributes may include creating calibration data; which would involve data generated in-house for test purposes so as to better understand field data and provide a benchmark for comparison during the modeling phase.

Considerations:

- Can missing attributes be reconstructed by aggregation, averaging, induction or extrapolation?
- Does the data need calibration or normalizing?
- Does data need to be modified in order to feed the modeling phase?

3.5 Integrating Data

Multiple datasets will need to be used during the project, such as the combination of existing as built data with newly acquired field or inspection data. It will be necessary to properly integrate these dataset in order to address business concerns. Data integration can be achieved in two (2) basic ways:

- Merging two or more datasets that concern the same rows but have different attributes, resulting in an increase in columns of the consolidated data table. An example would be applying electromagnetic and leak detection data to the same pipe distance (same number of rows/pipe sticks).
- Appending two or more datasets that concern the same attributes but refer to different objects. An example would be applying test or calibration data to field data in order to benchmark and achieve a better understanding of the latter.

Once data integration is complete, the data exploration phase may be revisited in order to ensure that the data integration task has been performed correctly.

Considerations:

- Data should be saved prior to outputting the results to modeling.
- Can data values be aggregated after integration? Can new values be computed to summarize table information?

3.6 Formatting Data

The final step of the Data Preparation phase is to ensure that the data is in the appropriate format to feed the Modeling phase. It is not uncommon that for example lay sheets may have measurements in either metric or imperial stationing (or both!) and it is critical that the data being output for modeling is consistent and in line with the modeling requirements.

Considerations:

- Does the data exist in the format or order required by the model?
- How does the model affect the datasets?

4. Modeling

4.1 Modeling Overview

The purpose of this phase is to develop an algorithm that fits the data – or, in other words, a data-driven technical solution in order to address the business problem. Thanks to this model, the company will also be able to predict the future, extract insights and innovate – all leading towards a profitable solution[13].

Prior to explaining the different modeling techniques, it is important to remember the initial goal – the problem to be solved – because this will help in navigating through the different models available. A business problem can be resolved in three ways: predicting, estimating, or identifying the quantity of specific attributes. All three of them entail discovering relationships and leveraging them to make effective business decisions, but the procedure from a pure data analysis perspective will differ.

A utility company, for example, wants to know what the condition of asset pipelines is, which pipes are healthy and which pipes are distressed and if are they going to fail. This will help to plan their capital expenditure, perform proactive maintenance and avoid catastrophic failures. One of the attributes to consider in this assessment process is the pipeline installation date.

Pipelines are classified according to manufacturing date and therefore constructed in accordance to different industry standards - typically in compliance with the American Water

[13] Peng and Matsui (2016) p.55

Works Association (AWWA). As a result, pipes manufactured at different points in time will vary in regards to the amount of distress they can sustain prior to reaching their yield point and failing. Consequently, pipe manufacturing date and specifications are data which must be included in the structural analysis model since this is computing the level of distress in order to predict failures and prioritize maintenance and capital expenditure.

This type of supervised learning is known as classification and takes a proactive role in leveraging information to facilitate prediction.

However, the analysis of data does not end here. Remember that it is essential that to remain in continuous communication with relevant stakeholders in order to ensure that the technological advances go hand in hand with the business aspect of the company.

4.2 Selecting Modeling Techniques

Following the above example, to consider the relationship between business goals such as condition assessment and proactive maintenance and the type of data available, such as manufacturing specifications, and relay their importance to business and managerial stakeholders. As the relationship may not be obvious or immediately apparent, it is possible to demonstrate the modeling requirements by grouping problematic pipelines together, while disregarding previous findings (manufacturing date/specifications/standards), and let the algorithm look for patterns through an analytical process known as clustering.

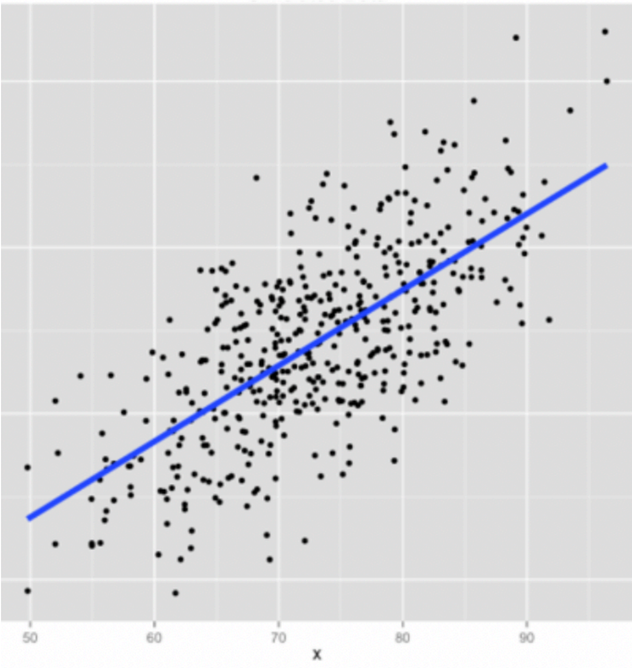


Figure 6: Example of what a cluster graph on Historical Pipeline Failures by material and date could look like?????

Such a model automatically studies the clusters until the algorithm finds that the pipelines that have suffered from failures or are suspect of leaks, were not only constructed before 1980 but also made up of prestressed concrete cylinder pipe (PCCP).

This type of analytical technique is a form of unsupervised learning known as clustering, which consists of enabling an algorithm to group observations and find common patterns in an automated manner.

To do this, the responsible person must create a simple regression analysis that indicates how many pipelines were constructed prior to 1980 using PCCP and illustrate this in the form of a histogram. As a result, business managers will have an accurate estimate of which asset pipelines are likely to require inspection or maintenance and in the following years and make informed capital budget planning estimates.

As such, use of data predictive data analytics will have transformed a question mark into informed estimate that adds business value.

Thus, it is possible to identify which pipelines are likely to fail in the future and why. This would certainly be valuable to the company, but from a business perspective it is more important to also know the condition that the pipeline is in.

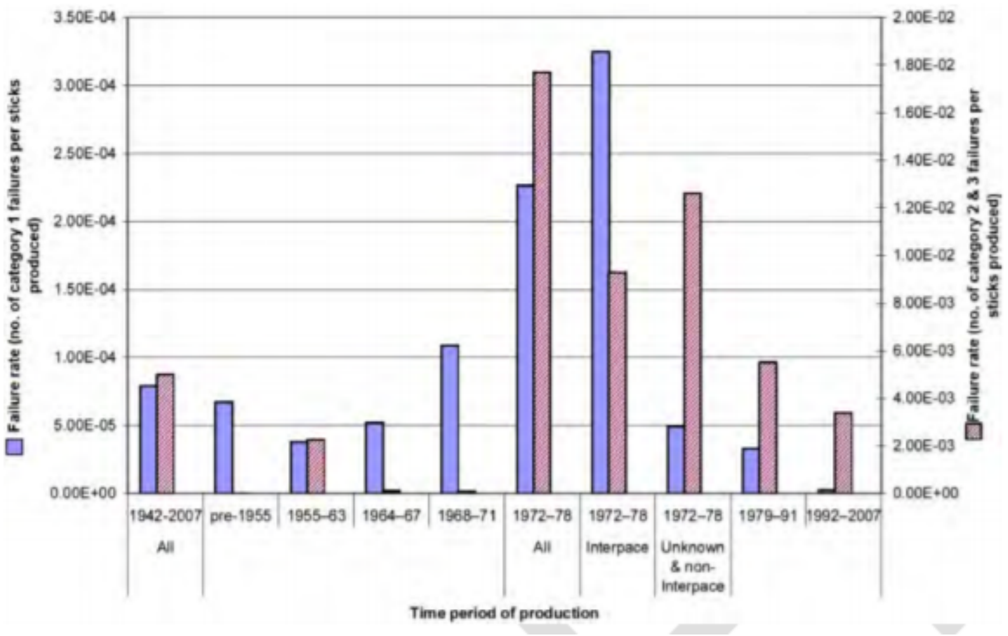


Figure 7: Failure of PCCP by Pipe Vintage

This is a good way of illustrating three main ways of deconstructing complex data:

- **Classification:** Consists of arranging observations according to predefined categories that enable the possibility of predicting outcomes in the future.
- **Regression:** By quantifying observations according to preselected variables it is possible to estimate the value of a future event.
- **Clustering:** For a newcomer to data science, clustering may seem a little similar to classification. Whilst the core is indeed the same – grouping observations –, clustering goes one step beyond. The process does not end once observations are grouped together. The objective is to then find commonalities, often referred to as patterns that either help in identifying or better understanding data relationships.

When deciding on the most appropriate model, consider the following:

The purpose of data analysis is to build a solution that is aligned with the predetermined business goals and adds value to the company: Data can help identify losses, expenditures, underperforming assets or services and can serve a multitude of purposes but, must ultimately contribute to adding value to the business. Specifically, identify leaks or distressed areas on the pipeline and developing distress trends that if left unattended may lead to catastrophic failures.

The data analyst, subject matter experts and business stakeholders must work together to understand what occurred in the past and what is occurring at present. An easy way for the data scientist to check if s/he has the whole picture is by answering the 5W questions: who, what, where, when and why. Yet, to go from a mere description of the situation (who, what, where) to a identifying the cause (why), the data scientist has to tick one last box: HOW. This alludes to the interrelationships between phenomena, which can only be understood if data science is combined with business acumen.

Current tools exist to understand the vast interrelationships between input layers, output layers and hidden layers, which will probably deliver a high predictive performance and numerous obscure patterns, which can be valuable in business decision-making[14].

Experts refer to this as the dilemma between White Boxes and Black Boxes. White Boxes are clear, simple, easy to explain and easy to understand. Black Boxes are a herculean task to decipher and comprehend. Guidotti, Monreale and Pedreschi claim that 'relying on sophisticated machine-learning classification models trained on massive datasets thanks to scalable, high-performance infrastructures, we risk to create and use decision systems that we do not really understand […] impacts not only information on ethics but also on accountability, on safety, and on industrial liability'[15]. Arguably, these collateral effects can result in multiple monetary losses. On the one hand, dissatisfied consumers ultimately lead to

[14] Dzeroski, S. (1996)

[15] Guidotti, Monreale and Pedreschi (2018) p.2

significant losses. On the other hand, miscellaneous costs, such as compensation packages or fines for pollution and environmental damage, can make a company go bust in a short matter of time.

The decision on which route to follow lies within the data team, business experts, managers and relevant stakeholders. It is possible that the budget and the success criteria established at the beginning provide some guidance when deciding whether to analyze a White Box or a Black Box. Nevertheless, regardless of what the company may decide, developing a model requires summarizing quantitative data and providing it with a structure[16].

4.3 Building the Models

Having examined the different approaches to complex analysis, it is time to tackle data science from a high-tech perspective, whether that is in the form of symbols or descriptive statistics or predictive mathematics with computations (i.e. machine learning).

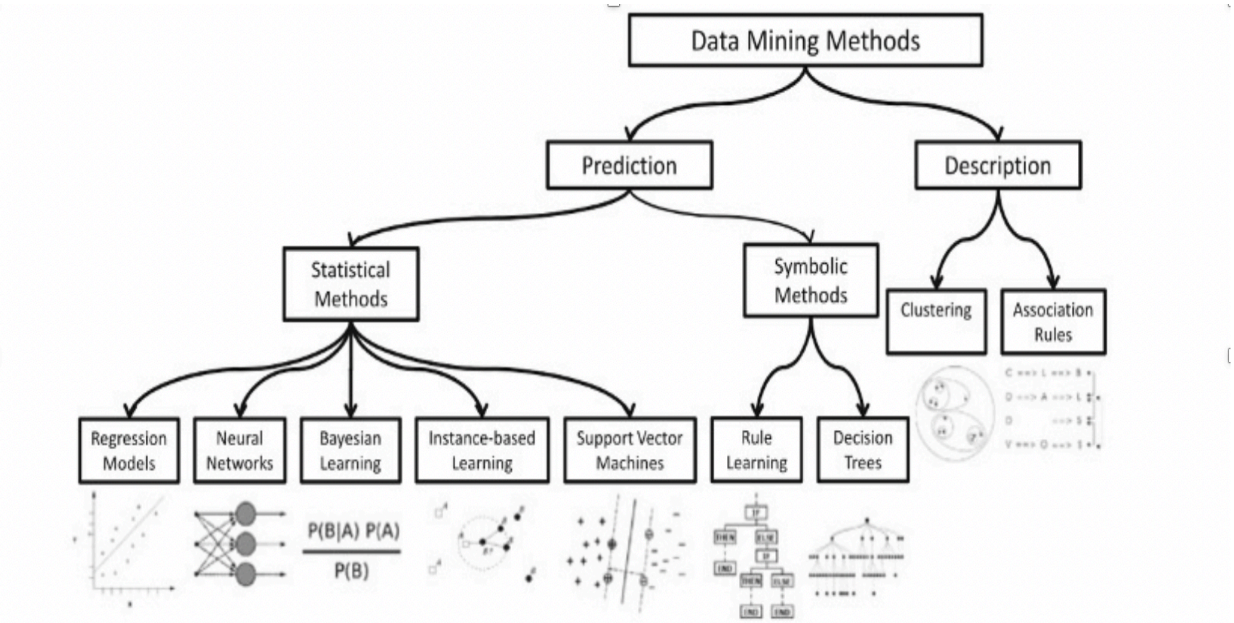


Figure 8: Data Mining Methodology Flow Chart[17]

[16] Peng and Matsui (2016) p.55
[17] García, Luengo, Herrera (2015) p.4

Symbolic methods, such as Rule Learning or Decision Trees are considered 'more interpretable for humans[18]'. Generally speaking, descriptive modelling does not intend to estimate or predict future events but rather 'gain insight into the underlying phenomenon or process[19]' as opposed to machine learning based on predictive mathematics. However, 'As machine learning methods were deployed broadly, the scientific disciplines of Machine Learning, Applied Statistics, and Pattern Recognition developed close ties, and the separation between the fields has blurred'.[20]

In order for the model to achieve the predetermined business goals, it is imperative to consider the following:

- What is the objective of data science in this business case? It is to describe a structural distress problem, estimate quantification of distress and predict future failures.
- What is the desired methodology to be used? Symbolic from a qualitative perspective, mathematical in terms of quantification of distress and statistical in terms of identifying clusters of distress.
- What are the stakeholders' business needs and/or desires? To identify distress in order to plan capital budgeting and prevent failures and service disruptions.

Through these questions, it will also become clear if the data must be managed as a White or Black Box – and consequently, supervised learning or unsupervised learning.

- **Supervised learning models** are designed to discover the relationships and associations between different variables (technically called, input attributes) and another variable previously selected as the object of study. Attributes range between unordered nominal attributes and logically ordered numeric attributes and they are normally studied through classification and regression. As the data scientist strives to establish connections between known attributes, supervised learning is normally used in White Box cases.

[18] García, Luengo, Herrera (2015) p.4
[19] Provost and Fawcett (2013) p.46
[20] Provost, Fawcett (2013) p.39

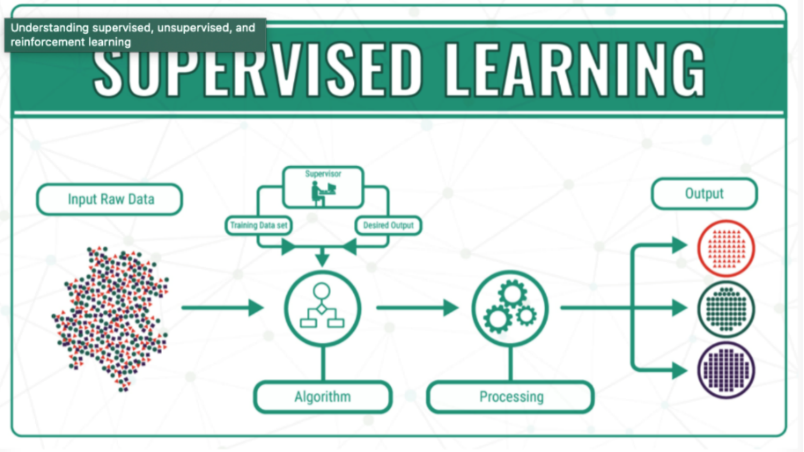


Figure 9: Supervised Learning Methodology Flow Chart[21]

- **Unsupervised learning models** are significantly more complex than supervised models. With unsupervised models, the data scientist will probably only compute two variables at a time. Unsupervised models aim to bring to light the numerous hidden interconnections between multiple attributes. As a result, it is possible to obtain numerous 'regularities, irregularities, relationships, similarities and associations in the input[22]' illustrating the deep vertical, horizontal and multidirectional hierarchies below the known surface. The most common modeling techniques for unsupervised learning are clustering and other associational rules.

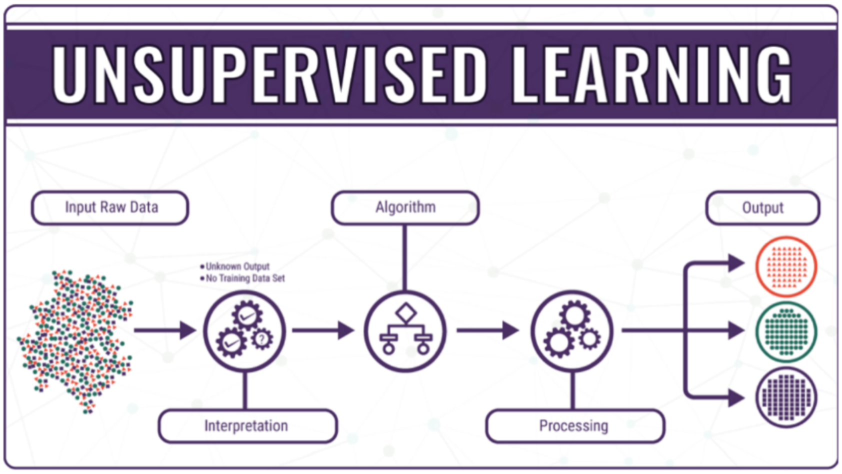


Figure 10: Unsupervised Learning Methodology Flow Chart[23]

Moreover, when designing the model, there are a few important business-oriented considerations:

[21] Chi Software. Supervised and Unsupervised Machine Learning. (2019)
[22] García, Luengo, Herrera (2015) p.7
[23] Chi Software. Supervised and Unsupervised Machine Learning. (2019)

- **Objective of whole discovery process**: The business goals should be at the center of decision-making. For instance, in this phase, the objective of the process can determine the breadth and depth required to compile the descriptions, predictions or estimations desired. In other words, find descriptive patterns or leverage these patterns to formulate predictions and estimations. It is the intended added value of the discovery process what will guide the data scientist in these decisions.
- **Maintenance**: The business objective also clarifies whether this discovery process unique in time or, contrarily, if the goal is to build a recurrent model that continuously feeds decision-making. In this case, the data scientist must design an algorithm that can be easily comprehended and maintained.
- **Budget**: As the ultimate goal of the knowledge discovery process is to add value to a company, this process must be completed in a manner that is cost-effective. To avoid complications in the latter stages, the data scientist must communicate concerns arising in this phase to relevant stakeholders in order to guarantee that the model adds business value to the company.

Thus, the purpose of the model design is to transform data into a valuable asset for the company. All companies produce data – but not all companies know how to use this data to create value for their business. This is the objective of the modeling phase. As such, numerous and different models may be tried until arriving to one that works best for the data obtained and the business objective.

In the case of predictive pipeline condition, the assessment of data would have to be combined in the following way:

- Consolidate existing pipeline information (lay sheets, plan and profile drawings and GPS coordinates) in order to build a pipe chart.
- Apply acoustic leak detection data, electromagnetic structural data and pressure data to the pipe chart in order to identify clusters of distress.
- Utilize pipe manufacturing specifications and pressure data in order to calculate the capacity of a given pipeline to sustain distress as well as its yield and strength limits.
- Review how the quantification of distress and clustering relates to these limits (is it below the limits or does it exceed them).

Once the modeling is complete it is important to reflect on the following:

- Can meaningful insights or conclusions be drawn? Are there trend patterns or developing clusters that can easily be identified?
- Was the processing time reasonable and within budgetary constraints?
- Did the calculations result in inconsistencies? Are there anomalies present? How do the parameters of the model affect the results?

4.4 Assessing the Model

4.4.1 Iteration

The first model designed, will certainly not be the final one. CRISP-DM difference from other methodologies is that it is an iterative model. This means that every phase can be revisited and reconsidered at any point in time when new information appears. This is especially pertinent in this phase.

By employing three different analytical techniques to convert data into business value. An initial classification led to finding out that PCCP built before 1980 had higher failure rates, the second analysis using clustering discovered that the cause of the problem was the material with which the pipelines were constructed, and the final analysis using a regression model identified how many asset pipelines would require immediate maintenance. The iteration in the analytical process was what eventually led to the business value.

Similarly, when computing an algorithm for the model, the data analysis process will rely on hypotheses and expected outcomes and key predictors to form associations between observations[24]. These preconceived factors should be accurately based on the business understanding and data-related phases which precede the modelling phase. Yet, just as with the example of vintage PCCP, the model may need to compute a variable that is either inaccurate or incorrect.

In order to understand if the model delivers the desired results, consider the following:

- Review the results based on the understanding of the business problem, consulting subject matter experts who may provide additional insights into the relevance of the particular results.
- Consider whether the results are deployable and if there are ways to narrow down on accuracy.
- Test the results by performing field validations, based on the model outputs.
- Analyze the results based on the success criteria established during the business understanding phase and determine whether the criteria have been met.
- Correct the model based on validations performed in order to benchmark future projects and improve future outputs.

4.4.2 Keeping track of revised parameters

Mistakes, errors, challenges or unwanted data frequently when appear when running the algorithm and it is highly unlikely that the solution to the business problem will not be optimal during the first assault.

[24] Peng and Matsui (2016) p. 96

As the Nobel Laureate Ronald Coase said, 'If you torture the data long enough, it will confess.[25]' These confessions will help in designing the best model possible for the extant data and the business problem, but not without having to deal with a couple of challenges, such as:

- **Overfitting** commonly occurs in machine learning-dependent unsupervised learning models, which attempt to identify associations with all kinds of available data, including that which is irrelevant and can potentially contaminate the model's findings. To overcome this, 'What we need to do is to "hold out" some data for which we know the value of the target variable, but which will not be used to build the model'[26]. For example, when creating and testing for a calibration model to benchmark the results of an electromagnetic inspection and output dataset, a number of parameters are constantly being tweaked. These parameters are gradual increase of electromagnetic frequency, current amplitude, electromagnetic wavelength, signal strength and receiver sensitivity, in order to obtain the most useful information out of the inspection data While increasing these parameters for example may initially result in an increase of useful signal up to a certain value point, further increase could result in signal saturation, thereby decreasing useful information captured or resulting in false positives.
- **Underfitting** occurs when the model cannot capture all the relevant data. As a result, the algorithm shows low variance but high bias.

To resolve these issues, the model may need to be adjusted as many times its iterations indicate flaws. However, note that a technically perfect model is yet to be designed. A business-oriented useful model, on the other hand, must be achieved to the uttermost extent since the goal of the model is to contribute value to the business.

To accomplish this, the team will have to revisit each stage and continuously reexamine the data computed in each phase and the knowledge extracted. It is only through constant iteration across all phases, feeding back on each decision with new information on the data relationships, that the team will succeed in designing a useful model. As CRISP-DM is a cyclical methodology it would be typical at this stage to not only iterate the model, but to also move back and forth between phases, and assess how the data and different decisions affect the model.

[25] Provost and Fawcett (2013), p.11.
[26] Provost and Fawcett (2013) p.113

5. Evaluation

5.1 Evaluation Overview

The evaluation phase assesses whether the business objectives were reached. Once the business needs are satisfied, the entire process will be reviewed in order to identify any aspects, which were overlooked. Ultimately, a decision should be made on how to proceed with results.[27]

5.2 Evaluating the Results

The evaluation phase essentially confirms that the analysis conducted during the modeling phase results in tangible outputs according to the data analysis success criteria, as well as the business success criteria. It is important to consider the following:

- Are the results clear and are they relatable to the business problems?
- Is it possible to identify data trends that point towards specific problems areas (distress or leak locations or hydraulic pressure issues)?
- Are there unexpected results or anomalies that do not make immediate sense, such as discrepancies between datasets or expected results that should be highlighted or analyzed further?
- Do the results provide any insights regarding confidence in our model and is it necessary to manage expectations?

5.3 Review Process

The evaluation phase provides us with the opportunity to reflect on the decisions and assumptions made during the entire process. In order to validate the results effectively it is important to consider the following:

- Have underlying assumptions, decisions and activities been properly documented and does their review provide insights towards the results? For example, are the distress patterns that we are seeing typical or do we run the risk of reading false positives? Are they comparable to patterns or data trends from previous projects? Are the pipe specifications that were fed into the model accurate? Or are they based on assumptions due to lack of available datasets and information? How can the assumptions or decisions made affect the results of the data analysis? If they can affect our output results, does the modeling process need to change?
- Are there other decisions that can be made during the data understanding, preparation and analysis process that could change the reading of the data results? For example, pipe specifications if inaccurate or missing (in which case they can be assumed based

[27] Crisp-DM Methodology Luchtvarfeiten.nl

on observation of external characteristics and installation date) can lead to using a different measurement prototype and quantification of distress can vary significantly.

- Do the results indicate any clear trend patterns of distress? Can the results be adequately quantified and accurately located?
- If the results do not indicate any obvious trend patterns (distress-related or otherwise) can we question our input or process and arrive at different conclusions?
- Are there possible ways to streamline our results? Can we use previous project experience to benchmark the results in a different way? Is there something that can be done in a better way or an internal process that can be improved? For example, can the process of data collection be improved in some way? Can the data analysis team improve efficiency during the modeling phase tasks?

5.4 Determining the Next Steps

Having completed data collection and analysis process, and after evaluating the results in respect to the goals and success criteria determined during the business understanding phase, it is necessary to decide how to further proceed.

- If the data analysis results are evaluated satisfactory, the final report can be produced.
- If the data analysis results are suboptimal then another iteration of the modeling phase should be taken into consideration, especially if a different set of decisions such as using a different prototype or benchmarking process, can lead to a different output result.

Figure 11: Data analysis results can be used to drive decision that involve significant financial impact

6. Deployment

6.1 Deployment Overview

The deployment phase considers the results of the evaluation to determine a strategy. It is important at this stage that the company is aware of the required actions to take. The final deliverable from this phase includes a final report and presentation of the obtained results.[28]

6.2 Planning for Deployment

The first step towards preparing for deployment of the data analysis results is planning for a clear and meaningful way to present them. Consider the following:

- Prepare a clear and comprehensible of both summary of the data analysis findings, as well as a summary of the methodology used to derive them in the modeling stage. It is critical that the results are justifiable and defensible with a transparent modeling process.
- Present the methodology used for each dataset and how it results in a conclusive finding. For example, when a conclusive finding is a leak or an area of distress, then it is important to explain how the data was collected, combined, modeled and analyzed and how the conclusion of the problem was reached. It is not necessary to delve deep into technical details in the main body of the report (a more technical analysis can be attached as an appendix) but it is important to be transparent about the limitations of the process in order to manage expectations.
- Identify the conclusive findings (these could be amount and location of distress, number and location of leaks, presence of pressure transients, structural capacity) and propose recommendations on how to address them.

6.3 Planning Monitoring and Maintenance

Once conclusive findings have been identified, it is important to determine whether the proposed recommendation to address them will be effective. Are there other solutions that can be considered? And if yes, are they cost effective and in line with the strategic business goals set at the beginning of the project? Can recommendations be prioritized in order to address high priority concerns now and plan for medium or low priority concerns in a future timeline? Are there continuous long-term monitoring solution appropriate to address the project findings or should the project be repeated at a designated interval?

Consider the following:

- Consult subject matter experts to prioritize findings and determine a timeline for implementation.
- Plan for validation of the results to confirm findings.

[28] Crisp-DM Methodology Luchtvarfeiten.nl

- Follow up on validation measurements to check how they actual findings align with the data analysis findings and ensure that they are fed back to the data analysis model in order to improve the process or in worse-case scenarios that the results prove to be of low accuracy or result in false positives the process (or model) is redesigned and limitations can be highlighted for future projects.

6.4 Producing a Final Report

The final report is composed by taking into consideration all of the intermediate reports and notes that have been created at all the different stages of the project life cycle up to now, in order to tie up loose ends and present a consolidated summary of the project as a whole, including business understanding, data understanding, data preparation, data analysis methodologies, conclusive findings and recommendations. It is important that the right information reaches the right people and is presented in a way which is understood by both decision makers, business personnel such as marketing and managerial staff, as well as technical experts who are responsible to carry out the maintenance tasks.

Specifically, the final report should include:

- A clear and concise description of the business problem.
- The data mining, data collection and data analysis methodologies and processes.
- Discussion on limitations and constraints regarding available data (or lack of) and how they impact the project and in what way.
- Discussion on deviations from the original scope (if any): Why they occurred, how they were handled and what they might mean for future projects?
- A summary of conclusive findings and an overview of recommendations to address findings, as well as roadmap for future steps forward.

6.5 Conducting a Final Project Review

The final project review is the last step in the CRISP-DM process. It is meant to reflect on the project within the context of an organizational discussion regarding what went well, what went wrong, what where the unexpected challenges and what can be improved in future projects. It is critical for the company to consolidate these discussions and build on lessons learned in a pursuit continuous improvement.

Appendix 1: Issues to consider before starting the process

1.1 Teams

In this comprehensive outlook of the tasks required to design a data-driven business solution to the problem of pipeline failures following the CRISP-DM methodology, there are two clear themes across the whole process:

- **The relationship between business-oriented and data-driven professionals is symbiotic.** In order to succeed on a business level – resolve a problem and avoid negative consequences – data is required. However, data needs to be understood within the framework of the business problem – and possibly, the grand scheme of the industry – in order to make data useful. Furthermore, whilst a company with a data-driven business model will have to ponder over how to convert data into a profit-making asset, a company with a non-data-driven business model (like the one which will be discussed in this handbook) must strive to utilize data to create a profitable solution.
- **As a Knowledge Discovery Process, the CRISP-DM methodology is iterative**. This means that as the team advances in the process, it will discover new knowledge, which the team should use to revisit and improve previous phases. Think of it as a form of feedback. In the modeling phase, specifically, the team will gather a vast amount of feedback on the decisions and assumptions made in the previous phases.

Thus, to succeed in a Knowledge Discovery Process, teams must necessarily count with data-focused people as well as business-oriented data experts. The latter, business-oriented data experts, are crucial since they will be the ones putting data into the "bigger picture" and translating data into business profitability. For these reasons, the leading figure is normally referred to as "business translator", a role commonly exercised by the Chief Analytics Officer/Chief Data Officer, whose vision is implemented by the business analysts should data analysts lack domain expertise[29].

On the other hand, data science teams can take different forms and titles depending on the company or the task. Essentially, what any company preparing to implement the CRISP-DM methodology should count with is a team that can collect and interpret data, navigate through Big Data, build a model, integrate and maintain the model in the day-to-day business operations, and produce data visualizations to ensure understanding[30].

Data teams can be contracted on a case-by-case level; in a pipeline company, however, data are an essential element to effectively conduct a pipeline company and will therefore already count with a departmental or cross-functional data-specialized team in the company.

Regardless of the structure, shape or form of the data science & analysis team, it is pivotal that it is aligned with the relevant business department(s) and the IT Department in order to develop a data-driven solution that fully accommodates the business needs and the company's

[29] Alexsoft. 2018. https://www.altexsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/
[30] Idem.

extant software. There are various ways to guarantee an effective integration, but we can venture in saying that all approaches follow a basic rule: 'There is nothing so useless as doing efficiently that which should not be done at all'[31]. A solution to the wrong problem is useless – and a good solution built with tools and resources (mainly, software) that the company does not possess also defeats the solution's effectiveness.

The iteration that characterizes the CRISP-DM methodology can help to avoid falling into the trap of developing a solution to the wrong problem or the wrong solution to the right problem[32]. Provided that the working group exchanges findings and conclusions in a synchronized manner as they advance through the Knowledge Discovery Process, the data-driven model can prove both relevant and useful in extracting useful knowledge from large databases[33].

1.2 Timeline

The first four phases should be completed promptly in order to carry out the first iteration in one week. When reading this handbook's discussion on those phases, it is likely that the reader thinks that these exhaustive phases can never be completed in such a short space of time.

It may help to recall that the purpose of the CRISP-DM is to resolve a particular company's specific business problem. Yet, if the reader examines this process critically, s/he will understand that when there is a business problem, there are actually two issues to deal with: The business problem and the business solution.

Developing a useful solution to the business problem will entail overcoming many challenges that will arise through the iterations, as is discussed throughout the handbook. At a glimpse, the first iteration will indicate, for example, if the data selected is relevant, available or useful. This is because even if the first few phases are completed thoroughly, neither the Chief Data Officer nor the business stakeholders will possess a complete understanding of the business problem until they acknowledge how the problem reacts to different possible solutions as this will enable the working group to sharpen the questions to be answered. Put simply, you will not know if a cake tastes good until you try – even if you have used all the right ingredients. Thus, this handbook recommends not delving into the first few phases for more than a week.

Furthermore, it is pertinent to advise the reader to prioritize usability and efficacy over perfectionism. It is likely that the team is never fully satisfied with the results, but the key

[31] The Effective Executive. Peter Drucker. 2016

[32] KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185. Archived January 9, 2013, at the Wayback Machine. Azevedo, A. and Santos, M. F.

[33] Fayyad, U. M., 1996. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, Vol. *11 No.* 5, pp 20-25.

question is whether the model is useful. Therefore, should a model leverage useful results after a number of iterations, it is best conclude that this model is apt as a business solution.

Bibliography

Introduction to Data Mining. Chapter 2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Pearson New International Edition, First Edition, 2013. ISBN: 978-1292026152

From Data Mining to Knowledge Discovery in Databases, Usama Fayyad, Gregory Piatetsky-Shapiro, From Data Mining to Knowledge Discovery in Databases and Padhraic Smyth, American Association for Artificial Intelligence, 1996

CRISP-DM Methodology: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185. Archived January 9, 2013, at the Wayback Machine. Azevedo, A. and Santos, M. F.

The Data Bonanza: Improving Knowledge Discovery for Science, Engineering and Business. Malcolm Atkinson, Rob Baxter, Peter Brezany, Oscar Corcho, Michelle Galea, Jano van Hemert, Mark Parsons, and David Snelling. John Wiley & Sons Ltd., 2013. ISBN: 978-1118398647

The Fourth Paradigm: Data-intensive Scientific Discovery. Tony Hey, Stewart Tansley and Kristin Tolle. Microsoft Research, 2009. ISBN: 978-0982544204

Data Preprocessing in Data Mining. Salvador García, Julián Luengo, Francisco Herrera. Springer-Verlag, 2015. ISBN: 978-3319102467

The Art of Data Science. A Guide for Anyone Who Works with Data. Roger D. Peng and Elizabeth Matsui, 2016.

Data Science for Business. Foster Provost and Tom Fawcett, 2013.

Riccardo Guidotti, Anna Monreale and Dino Pedreschi (KDDLab, ISTI-CNR Pisa and U. of Pisa). Available at: https://www.kdnuggets.com/2019/03/ai-black-box-explanation-problem.html

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (August 2018), 42 pages. DOI: https://doi.org/10.1145/3236009.

IBM Corporation, IBM SPSS Modeler CRISP-DM Guide

Crisp-DM Methodology, A Structured Approach for Data Mining Projects, Luchtvartfeitenn.nl

Business Analytics: A Framework, Rafi Ahmad Khal et al, International Journal of Computer Technology & Applications vol.10 (2), 102-108

DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model, Steffen Huber, Hajo Wiemer, Dorothea Schneider, Steffen Ihlenfeldt July 2018

Courtenay Pump Station 750/820-Millimetre Force Main Condition Assessment Report, Comox Valley Regional District, V. Sagiannos, June 2017

Pure Technologies, Innovative Solutions that Save Time, Money And Conserve Your Water Resources https://puretechltd.com/water-and-wastewater/

Pure Technologies, Pipeline Condition Assessment https://puretechltd.com/water-and-wastewater/solutions/pipeline-condition-assessment/

Xylem Inc., Xylem showcases advance infrastructure analytics solutions https://www.xylem.com/en-us/making-waves/water-utilities-news/xylem-showcases-advanced-infrastructure-analytics-solutions-in-singapore/

Dzeroski, S. 1996. Inductive Logic Programming for Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 59–82. Menlo Park, Calif.: AAAI Press.

Fayyad, U. M., 1996. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, Vol. *11 No.* 5, pp 20-25.

CHI Software. Supervised vs. Unsupervised Machine Learning. Available at: https://medium.com/@chisoftware/supervised-vs-unsupervised-machine-learning-7f26118d5ee6

Xylem Inc., The Rise of Digital Twins and Decision Intelligence in Water Utilities.

The Effective Executive. Peter Drucker. 2016