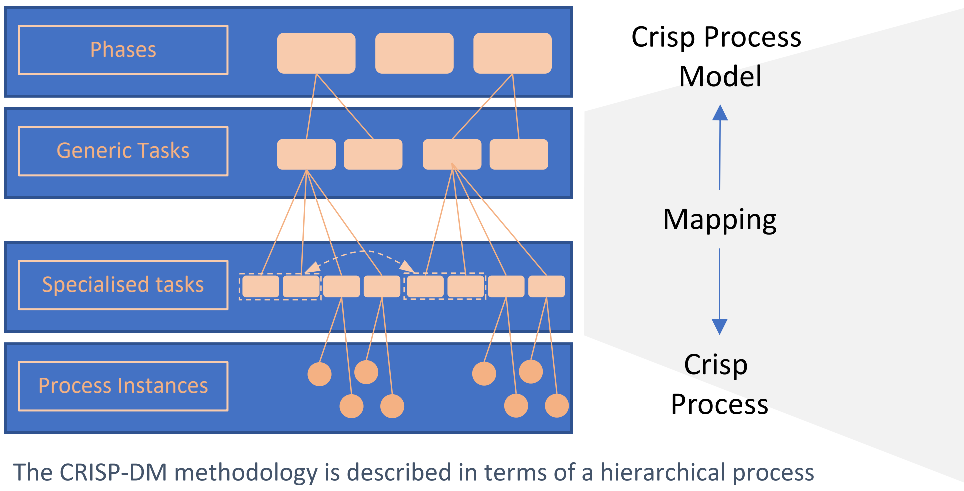
CRISP-DM methodology



	Data Mining Context			
Dimension	Application Domain	Data Mining Problem Type	Technical Aspect	Tool and Technique
Examples	Response Modeling	Description and Summarisation	Missing values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
		Concept Description		Decision Tree
		Classification		
		Prediction		
		Dependency Analysis		

The CRISP-DM methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance.

The description of phases and tasks as discrete steps performed in a specific order represents an idealized sequence of events. In practice, many of the tasks can be performed in a different order, and it will often be necessary to repeatedly backtrack to previous tasks and repeat certain actions. Our process model does not attempt to capture all of these possible routes through the data mining process because this would require an overly complex process model.

Mapping for the present

Generic process model mapped to a single data mining project ➤ ad-hoc/single use

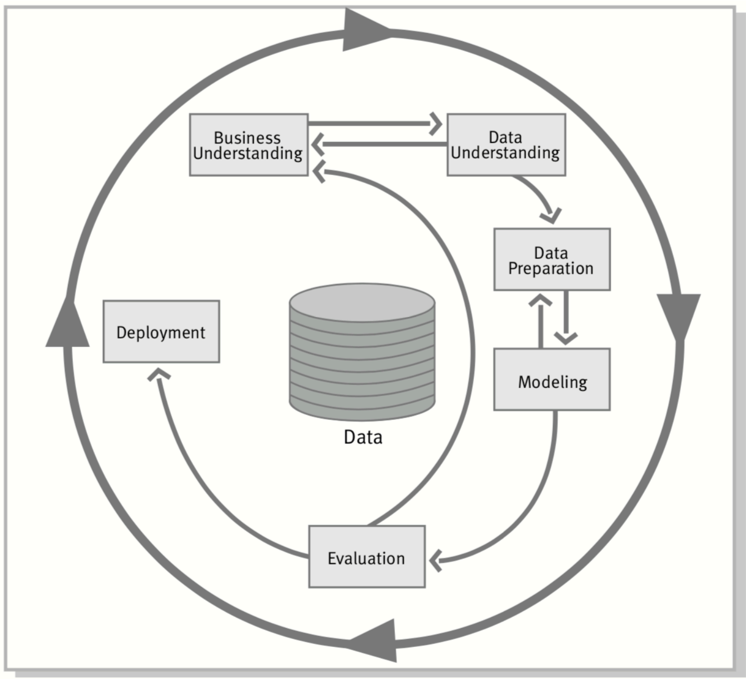
Mapping for the future

Systematic analysis and consolidation of experiences of a single project ➤ specialised process model, re-use

How to map the generic process model to the specialised level.

- Analyse your specific context
- Remove any details not applicable to your context
- Add any details specific to your context
- Specialize (or instantiate) generic contents according to concrete characteristics of your context
- Possibly rename generic contents to provide more explicit meanings in your context for the sake of clarity

Phases of the CRISP-DM reference model



Glossary/terminology

Activity – Part of a task in the User Guide; describes actions to perform a task

CRISP-DM methodology – The general term for all concepts developed and defined in CRISP-DM

Data mining context – A set of constraints and assumptions, such as problem type, techniques or tools, application domain

Data mining problem type – A class of typical data mining problems, such as data description and summarization, segmentation, concept descriptions, classification, prediction, dependency analysis

Generic – A task that holds across all possible data mining projects

Model – The ability to apply algorithms to a dataset to predict target attributes; executable

Output – The tangible result of performing a task

Phase – A term for the high-level part of the CRISP-DM process model; consists of related tasks

Process instance – A specific project described in terms of the process model

Process model – Defines the structure of data mining projects and provides guidance for their execution; consists of reference model and user guide

Reference model – Decomposition of data mining projects into phases, tasks, and outputs

Specialized – A task that makes specific assumptions in specific data mining contexts

Task – A series of activities to produce one or more outputs; part of a phase

User guide – Specific advice on how to perform data mining projects

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding

The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

Data preparation

The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modelling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modelling tools.

Modelling

In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

Evaluation

At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes—for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

Phase 1 - Business Understanding

Reference Model Phases | Tasks | Outputs | 'what'			User Guide Phases | Tasks | 'how'	
Task	Outputs	… to include		
Determine Business Objectives	Background What the customer really wants to achieve	Record the organisational situation at project start	*Organization* • Develop organizational charts; names and responsibilities • Stakeholder map; Identify the business units which are affected by the data mining project • Identify an internal sponsor (financial sponsor and primary user/domain expert) • If there is a steering committee and list members • Identify the business units which are affected by the data mining project (e.g., Marketing, Sales, Finance) *Current solution* • Describe any current solution currently • Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users	*Problem area* • Identify the business area where the problem exists, • Describe the problem in general terms Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business) Clarify prerequisites of the project (e.g., What is the motivation of the project? Does the business already use data mining?) If necessary, prepare presentations and present data mining to the business • Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?) • Identify the users' needs and expectations
	Business Objectives Getting to the crux of the detail Helps build knowledge	Primary business objective. Any other questions to address?	• Informally describe the problem to be solved • Specify all business questions as precisely as possible • Specify any other business requirements (e.g., the business does not want to lose any customers) • Specify expected benefits in business terms	*Beware of setting unattainable goals—make them as realistic as possible.*
	Business Success Criteria	Criteria needs to be measured objectively	• Specify business value-based success criteria • Identify who assesses the success criteria • Each of the success criteria should relate to at least one of the specified business objectives • Decide the evaluation strategy to be used | Include review points | End of Phase review, update next phase/rest of plan.	*Before starting the situation assessment, you might analyse previous experiences of this problem—* *either internally, using CRISP-DM, or externally, using pre-packaged solutions*
Assess Situation	Inventory of resources	Resources | Data | Computing resources | Software	*Hardware resources* • Identify the base hardware • Establish the availability of the base hardware for the data mining project • Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project • Identify the hardware available for the data mining tool to be used (if the tool is known at this stage) *Sources of data and knowledge* • Identify data sources (source type and format of data) • Identify knowledge sources (source type and format of data) • Check available tools and techniques • Describe the relevant background knowledge (informally or formally) Personnel sources • Identify project sponsor (if different from internal sponsor) • Identify technical people regarding data and knowledge sources • Identify market analysts, data mining experts, and statisticians, and check their availability • Check availability of domain experts for later phases	
	Requirements, Assumptions and Constraints	Schedule | Quality of Results | Security | Legal Issues | Data Approval Verifiable/non-verifiable Business Assumptions	*Requirements* • Specify target group profile • Capture all requirements on scheduling • Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability • of the data mining project and the resulting model(s) • Capture requirements on security, legal restrictions, privacy, reporting, and project schedule *Assumptions* • Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary) • List assumptions on data quality (e.g., accuracy, availability) • List assumptions on external factors (e.g., economic issues, competitive products, technical advances) • Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000) • List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor) *Constraints* • Check general constraints (e.g., legal issues, budget, timescales, and resources) • Check access rights to data sources (e.g., access restrictions, password required) • Check technical accessibility of data (operating systems, data management system, file or database format) • Check whether relevant knowledge is accessible • Check budget constraints (fixed costs, implementation costs, etc.)	
	Risks and Contingencies	Impact and mitigation on project schedule	*Identify Risks* • Identify business risks (e.g., competitor comes up with better results first) • Identify organizational risks (e.g., department requesting project doesn't have funding for the project) • Identify financial risks (e.g., further funding depends on initial data mining results) • Identify technical risks • Identify risks that depend on data and data sources (e.g., poor quality and coverage) *Develop contingency plans* • Determine conditions under which each risk may occur • Develop contingency plans	
	Terminology	Business glossary | Data mining terminology	• Check prior availability of glossaries; otherwise begin to draft glossaries • Talk to domain experts to understand their terminology • Become familiar with the business terminology	
	Costs and Benefits	Cost-benefit analysis	• Estimate costs for data collection • Estimate costs of developing and implementing a solution • Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue) • Estimate operating costs	*The comparison should be as specific as possible, as this enables a better business case to be made.* *Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.*
Determine Data Mining Goals	Data Mining Goals	Business goal | Data mining goal | intended outputs	• Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified). • Specify data mining problem type (e.g., classification, description, prediction, and clustering). For more details about data mining problem types, see Appendix	*It may be wise to re-define the problem. For example, modeling product retention rather than customer retention when targeting customer retention delivers results too late to affect the outcome*
	Data Mining Success Criteria	In technical terms | if subjective, record decision maker	• Specify criteria for model assessment (e.g., model accuracy, performance and complexity) • Define benchmarks for evaluation criteria • Specify criteria which address subjective assessment criteria (e.g., model explain ability and data and marketing insight provided by the model)	*Remember that the data mining success criteria are different than the business success criteria defined earlier.* *Remember it is wise to plan for deployment from the start of the project.*
Produce Project Plan	Project Plan	Stages | Duration | Resources | inputs-outputs | Dependencies | Identify iterations | relationship between risk and schedule | Detailed for each phase	• Define the initial process plan and discuss the feasibility with all involved personnel • Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria • Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating). • Identify critical steps | Mark decision points | Mark review points |Identify major iterations	
	Initial Assessment of Tools and Techniques	At end of phase	• Create a list of selection criteria for tools and techniques (or use an existing one if available) • Choose potential tools and techniques • Evaluate appropriateness of techniques • Review and prioritize applicable techniques according to the evaluation of alternative solutions	

Phase 2 - Data Understanding

Reference Model

Phases | Tasks | Outputs | 'what'

User Guide

Phases | Tasks | 'how'

Task	Outputs	… to include		
Collect Initial Data	Initial Data Collection Report	Acquiring, accessing data listed in resources Tool used Strategy for holding data Integration of multiple data sources/types Problems | Resolutions	*Data requirements planning* • Plan which information is needed (e.g., only for given attributes, or specific additional information) • Check if all the information needed (to solve the data mining goals) is actually available *Selection criteria* • Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?) • Select tables/files of interest | Select data within a table/file • Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise) *Insertion of Data* • If the data contain free text entries, do we need to encode them for modeling or do we want to group specific entries? • How can missing attributes be acquired? • How can we best extract the data?	*Be aware that data collected from different sources may give rise to quality problems when merged (e.g., address files merged with a customer database may show inconsistencies of format, invalidity of data, etc.).* *Remember that some knowledge about the data may be available from non-electronic sources (e.g., from people, printed text, etc.).* *Remember that it may be necessary to preprocess the data (time-series data, weighted averages, etc.)*
Describe Data	Data Description Report	Building a data dictionary: Format, quantity, identities of the fields, other surface details discovered	*Volumetric analysis of data* • Identify data and method of capture • Access data sources • Use statistical analyses if appropriate • Report tables and their relations • Check data volume, number of multiples, complexity • Note if the data contain free text entries *Attribute types and values* • Check accessibility and availability of attributes • Check attribute types (numeric, symbolic, taxonomy, etc.) • Check attribute value ranges • Analyze attribute correlations • Understand the meaning of each attribute and attribute value in business terms • For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.) • Analyze basic statistics and relate the results to their meaning in business terms • Decide if the attribute is relevant for the specific data mining goal • Determine if the attribute meaning is used consistently • Interview domain experts to obtain their opinion of attribute relevance • Decide if it is necessary to balance the data (based on the modeling techniques to be used) *Keys* • Analyze key relationships • Check amount of overlaps of key attribute values across tables *Review assumptions/goals* • Update list of assumptions, if necessary	
Explore Data	Data Exploration Report	Techniques to Query, Visualise and Report Relationships Aggregations Simple stats analysis Refinements, if any Feeder to transformations/data prep steps	*Organization* • Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations) • Identify characteristics of sub-populations *Form suppositions for future analysis* • Consider and evaluate information and findings in the data descriptions report • Form a hypothesis and identify actions • Transform the hypothesis into a data mining goal, if possible • Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable. • Perform basic analysis to verify the hypothesis	
Verify Data Quality	Data Quality Report	Is the data complete? Is it correct or does it contain errors? How are they represented, where do they occur and how common are they? List the results of the data quality verification; if there are quality problems, list possible solutions.	*Review keys, attributes* • Check coverage (e.g., whether all possible values are represented) • Check keys • Verify that the meanings of attributes and contained values fit together • Identify missing attributes and blank fields • Establish the meaning of missing data • Check for attributes with different values that have similar meanings (e.g., low fat, diet) • Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter) • Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon • Check for plausibility of values, (e.g., all fields having the same or nearly the same values) *Data quality in flat files* • If data are stored in flat files, check which delimiter is used and whether it is used consistently within all attributes • If data are stored in flat files, check the number of fields in each record to see if they coincide *Noise and inconsistencies between sources* • Check consistencies and redundancies between different sources • Plan for dealing with noise • Detect the type of noise and which attributes are affected	*Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).* *Use visualization plots, histograms, etc. to reveal inconsistencies in the data.* *Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behaviour (e.g., to check on customers' loan behaviour, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.).* *Review whether assumptions are valid or not, given the current information on data and business knowledge.*

General output for Phase 1 and 2

Phase 1 - Business Understanding

- Background
- Business Objectives and success criteria
- Inventory of resources
- Requirements, assumptions and constraints
- Risks and Contingencies
- Terminology
- Costs and Benefits
- Data mining goas and success criteria
- Project Plan
- Initial assessment of tools and techniques

Phase 2 – Data Understanding

- Initial data collection report
- Data description report
- Data exploration report
- Data quality report

Phase 3 - Data Preparation

Reference Model Phases | Tasks | Outputs | 'what'			User Guide Phases | Tasks | 'how'	
Task	Outputs	… to include		
Select Data	Rationale for inclusion / exclusion	List the data to be used/excluded and the reasons for these decisions (rows and columns)	• Collect appropriate additional data (from different sources—in-house as well as externally) Perform significance and correlation tests to decide if fields should be included • Reconsider Data Selection Criteria (See 2.1 Collect Initial Data) in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data) • Reconsider Data Selection Criteria (See 2.1 Collect Initial Data) in light of experience of modelling (i.e., model assessment may show that other datasets are needed) • Select different data subsets (e.g., different attributes, only data which meet certain conditions) • Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.) • Document the rationale for inclusion/exclusion • Check available techniques for sampling data	*Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types* *Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.*
Clean Data	Data Cleaning Report	Describe decisions and actions taken to address any data quality problems reported during the Verify Data Quality Task. Identify any outstanding data quality issues and what possible effect this could have on the results.	• Reconsider how to deal with any observed type of noise • Correct, remove, or ignore noise • Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer. • Reconsider Data Selection Criteria (See 2.1 Collect Initial Data) in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).	*Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.*
Construct Data	Derived Attributes	Attributes constructed from one or more existing attributes in the same record. Describe the necessity: representation of background knowledge, modeling use…	*Derived attributes* • Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate) • Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization) • How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).] • Add new attributes to the accessed data *Single-attribute transformations* • Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute) • Perform transformation steps	*Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps "income per person" is a better/easier attribute to use than "income per household." Do not derive attributes simply to reduce the number of input attributes.* *Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require them.*
	Generated Records	Describe creation of new records	• Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).	
Integrate Data	Merged Report	Merging of different tables. Also covers aggregations	• Check if integration facilities are able to integrate the input sources as required • Integrate sources and store results • Reconsider Data Selection Criteria (See 2.1 Collect Initial Data) in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)	*Remember that some knowledge may be contained in non-electronic format.*
Format Data	Reformatted data	Aligning the data with the requirements of the chosen modelling tool. Modifying the data where is loses no meaning	*Rearranging attributes* • Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict. *Reordering records* • It might be important to change the order of the records in the dataset. Perhaps the modelling tool requires that the records be sorted according to the value of the outcome attribute. *Reformatted within-value* • These are purely syntactic changes made to satisfy the requirements of the specific modelling tool • Reconsider Data Selection Criteria (See 2.1 Collect Initial Data) in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)	

General output for Phase 3

Phase 3 – Data Preparation

- Dataset description report (after pre-processing)
- Background, broad goals and plan for pre-processing
- Rationale for inclusion/exclusion of datasets

Phase 4 - Modelling

Reference Model — Phases | Tasks | Outputs | 'what'

User Guide — Phases | Tasks | 'how'

Task	Outputs	… to include		
Select Modelling Technique	Modelling Technique	The detail of what's to be used.	• Decide on appropriate technique for exercise, bearing in mind the tool selected.	*As the first step in modelling, select the actual initial modelling technique. If multiple techniques are to be applied, perform this task separately for each technique.*
	Modelling Assumptions	Record assumptions about the modelling technique and the data to be used.	• Define any built-in assumptions made by the technique about the data (e.g., quality, format, distribution) • Compare these assumptions with those in the Data Description Report • Make sure that these assumptions hold and go back to the Data Preparation Phase, if necessary	
Generate Test Design	Test Design	Describe the intended plan for training, testing, and evaluating the models.	• Check existing test designs for each data mining goal separately • Decide on necessary steps (number of iterations, number of folds, etc.) • Prepare data required for test	*Describe strategy for splitting the dataset into training, validation and test datasets.*
Build Model	Parameter Settings	List parameters and chosen values with their rationale.	• Set initial parameters • Document reasons for choosing those values	
	Models	Actual models, e.g. pkl files.	• Run the selected technique on the input dataset to produce the model • Post-process data mining results (e.g., edit rules, display trees)	
	Model Description	Model descriptions, expected accuracy, shortcomings, interpretation and any difficulties with their meanings.	• Describe any characteristics of the current model that may be useful for the future • Record parameter settings used to produce the model • Give a detailed description of the model and any special features • For rule-based models, list the rules produced, plus any assessment of per-rule or overall model • accuracy and coverage • For opaque models, list any technical information about the model (such as neural network topology) and any behavioural descriptions produced by the modelling process (such as accuracy or sensitivity) • Describe the model's behaviour and interpretation • State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts • about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)	
Assess model	Model Assessment	Interpretation of the model against domain knowledge, data mining success criteria and desired test design. Only evaluates the model. Results, rank and individual model quality.	• Evaluate results with respect to evaluation criteria • Test result according to a test strategy (e.g.: Train and Test, Cross-validation, bootstrapping, etc.) • Compare evaluation results and interpretation • Create ranking of results with respect to success and evaluation criteria • Select best models • Interpret results in business terms (as far as possible at this stage) • Get comments on models by domain or data experts • Check plausibility of model	*"Lift Tables" and "Gain Tables" can be constructed to determine how well the model is predicting.*
	Revised parameter settings	Hyperparameter tuning and further iterations	• Adjust parameters to produce better models.	

Phase 5 - Evaluation

Reference Model — Phases | Tasks | Outputs | 'what'

User Guide — Phases | Tasks | 'how'

Task	Outputs	… to include		
Evaluate Results	Assessment of Data Mining Results	Summarize assessment results in terms of business success criteria, including a final statement related to whether the project already meets the initial business objectives.	• Understand the data mining results • Interpret the results in terms of the application • Check effect on for data mining goal • Check the data mining result against the given knowledge base to see if the discovered information is novel and useful • Evaluate and assess results with respect to business success criteria (i.e., has the project achieved the original Business Objectives) • Compare evaluation results and interpretation • Rank results with respect to business success criteria • Check effect of result on initial application goal • Determine if there are new business objectives to be addressed later in the project, or in new projects • State recommendations for future data mining projects	*This step assesses the degree to which the model meets the business objectives, and seeks to determine if there is some business reason why this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit.* *Moreover, evaluation also assesses other generated data mining results. Data mining results cover models that are related to the original business objectives and all other findings. Some are related to the original business objectives while others might unveil additional challenges, information, or hints for future directions.*
	Approved Models	Model results chosen based on meeting selected criteria	• After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.	
Review Process	Review of Process	Summarize the process review and list activities that have been missed and/or should be repeated.	• Provide an overview of the data mining process used • Analyse the data mining process. For each stage of the process ask: • Was it necessary? • Was it executed optimally? • In what ways could it be improved? • Identify failures • Identify misleading steps • Identify possible alternative actions and/or unexpected paths in the process • Review data mining results with respect to business success criteria	*At this point, the resulting model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the data mining exercise, the Process Review takes the form of a Quality Assurance Review.*
Determine Next Steps	List of possible actions	List possible further actions along with the reasons for and against each option	• Analyse the potential for deployment of each result • Estimate potential for improvement of current process • Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available) • Recommend alternative continuations • Refine process plan	*Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.*
	Decision	Describe the decisions made, along with the rationale for them.	• Rank the possible actions • Select one of the possible actions • Document reasons for the choice	

Phase 6 - Deployment

Reference Model
Phases | Tasks | Outputs | 'what'

User Guide
Phases | Tasks | 'how'

Task	Outputs	… to include		
Plan Deployment	Deployment Plan	Summarize the deployment strategy, including necessary steps and how to perform them	• Summarize deployable results • Develop and evaluate alternative plans for deployment • Decide for each distinct knowledge or information result • Determine how knowledge or information will be propagated to users • Decide how the use of the result will be monitored and its benefits measured (where applicable) • Decide for each deployable model or software result • Establish how the model or software result will be deployed within the organization's systems • Determine how its use will be monitored and its benefits measured (where applicable) • Identify possible problems during deployment (pitfalls to be avoided)	
Plan Monitoring and Maintenance	Monitoring and Maintenance Plan	Summarize monitoring and maintenance strategy, including necessary steps and how to perform them	• Check for dynamic aspects (i.e., what things could change in the environment?) • Decide how accuracy will be monitored • Determine when the data mining result or model should not be used any more. Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could no longer be used. (update model, set up new data mining project, etc.). • Will the business objectives of the use of the model change over time? Fully document the initial problem the model was attempting to solve. • Develop monitoring and maintenance plan.	*Monitoring and maintenance are important issues if the data mining results become part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan for monitoring and maintenance. This plan takes into account the specific type of deployment.*
Produce Final Report	Final Report	Bring all the results together, process, costs incurred, deviations, implementation plan and future work recommendations.	• Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models, etc.) • Analyse how well initial data mining goals have been met • Identify target groups for report • Outline structure and contents of report(s) • Select findings to be included in the reports • Write a report	*At the end of the project, the project team writes up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experience, or a final presentation of the data mining result(s).*
	Final Presentation	Summary use of Final report content.	• Decide on target group for the final presentation and determine if they will already have received the final report • Select which items from the final report should be included in final presentation	*As well as a final report, it may be necessary to make a final presentation to summarize the project— maybe to the management sponsor, for example. The presentation normally contains a subset of the information contained in the final report, structured in a different way.*
Review Project	Experience Documentation	Summarize important experience gained during the project. What well well. Improvements.	• Interview all significant people involved in the project and ask them about their experience during the project • If end users in the business work with the data mining result(s), interview them: Are they satisfied? What could have been done better? Do they need additional support? • Summarize feedback and write the experience documentation • Analyse the process (things that worked well, mistakes made, lessons learned, etc.) • Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?) • Generalize from the details to make the experience useful for future projects	

General output for Phase 4, 5, 6

Phase 4 – Modelling

- Modelling assumptions
- Test Design
- Model Description
- Model assessment

Phase 5 – Evaluation

- Assessment of data mining result with respect to business success criteria
- Review of process
- List of possible actions

Phase 6 – Deployment

- Deployment plan
- Monitoring and maintenance plan
- Final report

Appendix – Data Mining Problem Types

Types	Notes	Appropriate Techniques
Data Description and Summarisation	• Concise description of characteristics of the data, typically in elementary and aggregated form • Can be an objective in its own right • Applicable at the early stages • With exploratory data analysis, can provide first insights • Occurs in combination with other data mining problem types • Summarization also plays an important role in the presentation of final results	
Segmentation	• Manual or semi-automatic • Separation of data into interesting and meaningful subgroups • Could be an objective in its own right • Often segmentation is a means to solving other problems making data more manageable	• Clustering techniques • Neural networks • Visualization
Concept Descriptions	• An understandable description of concepts or classes • e.g. building business logic behind the segmentation and classification of the data • May not be complete, but covers the important groups • Purpose: to gain insights	• Rule induction methods • Conceptual clustering
Classification	• Assigning the correct class label to unseen and unlabelled data • An object that is discretely labelled characterised by features of different classes • Used for predictive modelling • Can be derived by segmentation before training a model • Connection to dependency analysis – between attributes • Note: analyse deviations/outliers before model building	• Discriminant analysis • Rule induction methods • Decision tree learning • Neural networks • K nearest neighbour • Case-based reasoning • Genetic algorithms
Prediction	• Similar to classification, but target is a continuous attribute • Examples in Regression and Time Series forecasting	• Regression analysis • Regression trees • Neural networks • K nearest neighbour • Box-Jenkins methods (forecasting) • Genetic algorithms
Dependency Analysis	• Describes significant dependencies (or associations) between data items or events • Mostly used for understanding rather than predictive modelling • Can be strict or probabilistic • Special type: Associations, Sequential patterns • Algorithm selection can be a challenge	• Correlation analysis • Regression analysis • Association rules • Bayesian networks • Inductive logic programming • Visualization techniques

In applications, dependency analysis often co-occurs with segmentation. In large datasets, dependencies are seldom significant because many influences overlay each other. In such cases, it is advisable to perform a dependency analysis on more homogeneous segments of the data.

Summary of dependencies

The following table summarizes the main inputs to the deliverables. This does not mean that only the inputs listed should be considered—for example, the business objectives should be pervasive to all deliverables. However, the deliverables should address specific issues raised by their inputs.

Refers to ⤴ Closely related to ↵	Background	Business Objectives	Business Success Criteria	Costs & Benefits	Data Mining Goals	Data Mining Success Criteria	Inventory of Resources	Project Plan	Requirements, Assumptions & Constraints	Risks & Contingencies	Terminology	Initial Data Collection Report	Data Description Report	Data Quality Report	Exploratory Analysis Report	Test Design	Models	Parameter settings	Assessment w.r.t Business Success Criteria	Review of Process	Deployment Plan	Maintenance Plan
Background																						
Business Objectives	⤴										↵											
Business Success Criteria		⤴																				
Costs & Benefits		⤴						↵														
Data Mining Goals		⤴							⤴													
Data Mining Success Criteria			⤴		⤴				⤴													
Inventory of Resources																						
Project Plan		⤴		↵			⤴		⤴	⤴												
Requirements, Assumptions & Constraints		⤴																				
Risks & Contingencies		⤴	⤴																			
Terminology	⤴	↵																				
Initial Data Collection Report					⤴		⤴															
Data Description Report	⤴													↵								
Data Quality Report												⤴	↵									
Exploratory Analysis Report												⤴										
Dataset & Dataset Description					⤴								⤴	⤴	⤴							
Test Design					⤴	⤴																
Models					⤴													↵				
Parameter settings					⤴												↵					
Model Description																⤴	⤴	⤴				
Assessment						⤴										⤴	⤴					
Assessment w.r.t Business Success Criteria			⤴								⤴											
Review of Process																			⤴			
Next Steps								⤴											⤴			
Deployment Plan									⤴													↵
Maintenance Plan									⤴												↵	
Final Report & Presentation											⤴								⤴			
Experience Documentation								⤴												⤴		