DOI: 10.1561/9781638282372.ch25

Chapter 25

Data Processing in Healthcare Using CRISP

By Roland Roller, Anne Marie Weggelaar-Jansen, Ricard Martínez Martínez, Sven Schmeier, Holmer Hemsen and Supriyo Chatterjea

25.1 The BigMedilytics Blueprint

The BigMedilytics Blueprint is based on the experience and lessons learned from the 12 BigMedilytics studies. It aligns them to an abstract level of common blocks based on their similarities. Big Data and AI are data-driven techniques that have many aspects in common with data mining, so instead of creating a new process model, we build our blueprint upon the cross-industry standard process for data mining, also known as CRISP-DM [1]. We use this well-established open standard process model, apply it to our healthcare domain, and incorporate our experiences and outcomes.

The CRISP-DM model is presented in Figure 25.1 and defines six phases, which will be described in the following paragraphs. As the figure indicates, CRISP-DM is not necessarily a fixed sequence of phases, as you can move back and forth if necessary. Moreover, the arrows connecting the different phases are not the only way to proceed. The outer circle shows a clockwise movement and highlights that the development of such a project is an iterative process. For instance, with the successful deployment and the completion of a project, previous problems or new ideas could be brought into a new use case, and the circle can restart. Also, the general process may continue after the end of the project.

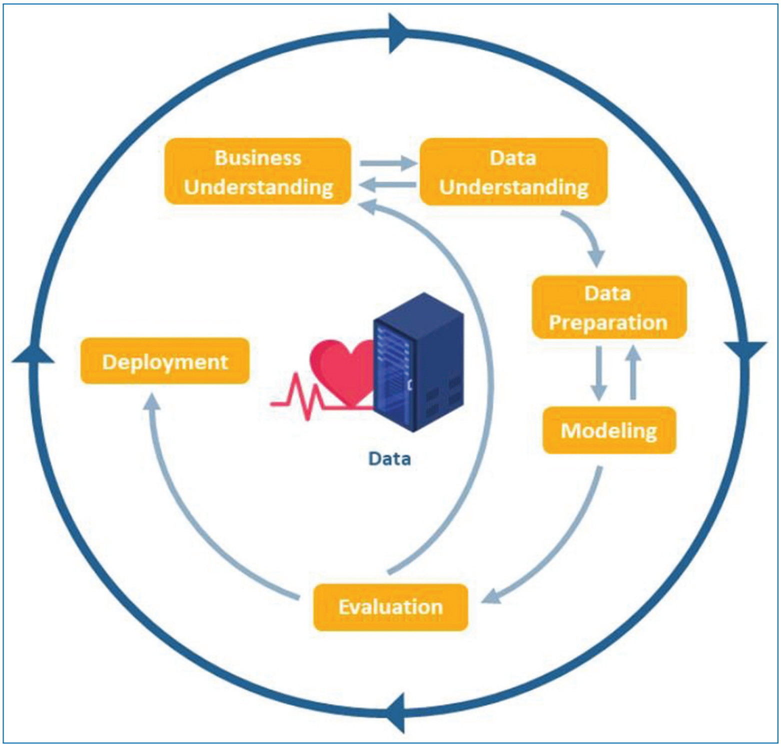

Figure 25.1. CRISP-DM—cross-industry standard process for data mining.

25.2 Business Understanding

An essential starting point in establishing a Big Data/AI project is a thorough understanding of the business, and the business goal(s), assessing the current and target situation, defining the technical task (Big Data/AI), and producing a project plan. Usually, this step results in a business model that describes how an organization creates, delivers, and captures value with the introduction of Big Data/AI technology.

Also, concerning business understanding, the healthcare sector behaves slightly differently. In healthcare, not all stakeholders are 'business' stakeholders, and a translation to our specific context is needed. Public, private for-profit, and private non-profit actors join forces to create value in the healthcare system. In addition, the value for patients is at least partly not only of monetary value. Still, it can be expressed in terms of quality of life, clinical outcomes, patient experience, and cost of treatment. Furthermore, value is also created for other stakeholders, such as healthcare professionals (better and faster decision-making and more efficient work processes), healthcare providers (higher productivity and better use of resources), and healthcare payers (better outcomes for the lowest cost). Also, value for the society can be discerned, e.g., a healthier population, increased labor productivity, and lower health expenses (in total or as a percentage of the GDP).

In BigMedilytics, all 12 studies developed a business model to describe what value their Big Data/AI innovation will create, which stakeholders, resources, and partners are involved or affected, and what activities are needed to create that value. Moreover, the business modeling process helps estimate which development costs and operational costs are related to the innovation and how such costs may be covered to ensure that the innovation action creates a positive value for the healthcare sector.

The classical 'business model canvas', a strategic management tool for developing and documenting new and existing business models, was part of BigMedilytics adapted for the context of Big Data/AI innovation in healthcare. The three main adaptations were as follows:

- Acknowledge the multi-sided market in healthcare and recognize that these innovations create value not only for patients but also for healthcare professionals, healthcare provider organizations, healthcare payers, and society at large.
- Acknowledge that 'profit' is not the main driver for innovating in healthcare. Still, it is a positive value in terms of better outcomes and/or lower costs.
- Acknowledge that, in the Big Data/AI context, rules and regulations play a key role and must be added as a separate player in the business model canvas.

Business modeling is not a one-time activity but often entails updates of the business model to finally meet the demands of all stakeholders involved in the activity. Therefore, it is rather a business modeling journey, moving from a business model of the pilot stage to the stage of scaling up and on to the phase of sustenance of the innovation.

In addition to that, compliance must be considered in this CRISP-DM phase as well. Therefore, the following requirements should be taken into account:

1. Integrate the Data Protection Officer and/or the Compliance Officer in the design team early.
2. Bear in mind that not only European legislation (GDPR) must be taken into account but also the following must be considered:
   - National laws regarding research in health
   - Laws of other member states in trans-European projects
   - Laws associated with international data transfers
   - 'Soft Laws' (such as Guidelines from the European Data Protection Board (EDPB), national data protection authorities, and so on).

3. In the case of multi-partner projects, roles should be defined in terms of GDPR (controller–joint controller–processor) and data use or access (data provider–data consumer). These roles can then be used to determine future agreements between the parties (joint-controller's agreement, processor's agreement, and data-sharing agreement).
4. Access to health system data requires considerations of ethical procedures(research protocol, ethical protocol, informed (ethical) consent, and ethics committee approval) as well as whether or not a clinical trial process of a medical device is necessary under Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.
5. Integrate the recommendations of the EU High-Level Expert Group on Artificial Intelligence, or depending on the location (e.g., the United States), alternatively the ethical principles of the Organization for Economic Cooperation and Development (OECD) and the ethical standards developed in the United States.

25.3 Data Understanding

The next phase is data understanding, which involves the sighting of the available data and the data collection. The data have to be explored and examined, particular characteristics identified, and the quality of the data assessed. This analysis needs to ensure that the data can help tackle the problem. If not, the aims within the business understanding must be amended, or other data sources must be found or incorporated.

Particularly in the medical domain, this phase can be challenging: Getting access to data can be time-consuming, for example, in case the data scientist is not associated with the data provider or is located in a different country. Also bear in mind that access to data requires having been able to comply with regulations in a way that allows you to demonstrate:

- The legitimate origin of the data
- The patient's consent, where necessary, and the guarantee of his or her rights
- The different requirements that each country imposes on retrospective and prospective studies.

In addition, data typically cannot be easily understood without medical (or additional) expertise from others. Therefore, interdisciplinary work is essential, which typically requires extra time, as different stakeholders use different terminology and understand the problem differently.

Furthermore, already in the data understanding phase, the problem of anonymization emerges. The position of the data protection authorities of the EU in Working Party Opinion 5/2014 on anonymization techniques is clear: irreversible anonymization must be achieved. Afterward, each data protection authority usually publishes its own guidelines.

The project has taught several lessons:

- De-identification and anonymization are not synonyms.
- Resources must be foreseen to verify the risk of re-identification and to perform second anonymization when health system data have been 'deidentified.'
- A double layer of additional measures should be implemented:
    - Technical: a controlled platform environment with appropriate security measures should be designed. Among these, the traceability of users is particularly relevant.
    - Legal: data-sharing agreements and non-identification commitments of partners and/or users of the platform should be formalized.

25.4 Data Preparation

In the next phase, the data have to be selected and integrated. Then, based on the previous analysis, data have to be cleaned and often converted into a different format to be used in the next stage. However, although data preparation and understanding might sound trivial, these steps usually take up most of the time within the overall project.

Often, in the healthcare domain, the creation and selection of digital data, such as electronic medical records, has grown over time, and so the quality, even of a single data source, may vary over time. However, missing and/or wrong information in the data is a common phenomenon that must be dealt with. Possible errors and inconsistencies might not be directly obvious to an outsider, emphasizing the need for interdisciplinary work on data cleansing. Moreover, valuable information might be "hidden" in not only structured but also unstructured text data, which makes it necessary to apply additional Natural Language Processing (NLP) techniques to access these data, which can raise additional challenges as those techniques need to be often adapted to the language and particular domain.

The limitation to working with the data only on-site, within a secure environment of the data provider, may also create additional challenges that need to be considered (e.g., limited user rights and old infrastructure).

25.5 Modeling

The next phase in CRISP-DM describes the modeling phase. In this phase, it needs to be decided which AI algorithms to use and which test procedure to follow. For example, we may choose to train a classifier based on the mathematical concept of linear regression and use a 10-fold cross-validation training strategy, i.e., we will perform the fitting procedure ten times, each time a training set of 90% of the data is randomly selected, keeping 10% of the data for validation.

Regarding Big Data, we also need to decide if we follow a batch (the data are collected and stored first and then in one or more batches analyzed) or a streaming approach (data are generated and analyzed continuously). In this respect, and especially with AI, a combination of both approaches is possible. For instance, a model may be trained in a batch but used with streaming data.

One of those lessons our 12 studies taught us is that due to the rather specialized data, the evaluation of the result of model training can only be interpreted successfully by a team of medical experts and data scientists.

As CRISP-DM focuses on data mining, there are certain aspects of modeling that, from a software engineering part, need to be taken into account concerning modeling but that are not addressed by CRISP-DM, such as modeling of the software architecture or domain modeling that may need to be considered in a healthcare project. Modeling involves the application of data protection by design principle. Usually, it starts with the data protection impact assessment from which the risks to be eliminated, mitigated, and reduced will be derived. Lessons learned from the project in this area include the following:

- The emergence of lists of criteria and methodologies of data protection authorities are similar but different. For example, the French National Commission on Informatics and Liberty (CNIL) offers a multilingual tool with a very open methodology. At the same time, the Spanish Spanish Data Protection Agency (AEPD) defines a comprehensive checklist of controls with a complementary guide of controls for AI.
- It is essential to train the whole staff.
- It is essential to consider the requirements of Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.
- Decisions must be taken on developing an ethical impact analysis on AI.

25.6 Evaluation

In this phase, an evaluation is performed if the previously defined business success criteria are fulfilled and supported by applying Big Data and AI technology.

If this is not the case, a thorough analysis of all previous phases is required. The reasons for not reaching the final goals may be manifold. For example, insufficient data are available to successfully train a model, and too much missing information or relevant data have not been collected. When all fails, it may even be necessary to amend the business success criteria in the business understanding phase. Moreover, this phase will review the overall work (e.g., were all steps properly executed? Summarize findings, and so on) and determine the next steps.

To measure the impact of Big Data innovations, business and research projects frequently rely on Key Performance Indicators (KPIs). In BigMedilytics, the three core KPIs were as follows: (a) improving the quality of the healthcare system by improving its effectiveness, (b) providing more people access to the healthcare system, and at the same time (c) reducing the costs by applying Big Data and AI technology.

But which KPIs capture the impact of Big Data innovations on healthcare in general? In the BigMedilytics project, we learned the following:

- First-order KPIs outline how Big Data innovations change the information provided.
- Second-order KPIs shed light on how Big Data innovations change the decision-making process.
- Third-order KPIs capture the perceived usefulness of the data and how far value is attributed to the information.
- And fourth-order KPIs reveal whether Big Data innovations might affect patient experience, population health, costs, and professional satisfaction in the long run.

However, KPIs might differ not only in their order effects but also in their function. Some KPIs can have a temporary function and can be revoked and revised as one sees fit. However, core KPIs remain relevant over time. Consequently, KPIs are an iterative, recursive process of moving back and forth between finding out which indicators are feasible, acceptable, measurable, and informative.

While assessing how Big Data affects long-term health outcomes, one needs to remember that long-term outcomes depend on a sequence of decisions and exogenous factors. How far changes in long-term outcomes can causally be attributed to Big Data innovations is therefore dependent on how rigorously one can establish the counterfactual scenario of what would have happened if the Big Data innovation had not been developed and implemented.

For evaluating Big Data and AI in healthcare, we also should keep in mind that the impact of applying the trained models for healthcare often can only be validated with new patients, i.e., over a longer period.

25.7 Deployment

The final phase defines the deployment of the model. This includes planning, roll-out, monitoring, and maintenance. Planning of the deployment should already be considered in the business understanding phase as the deployment of the Big Data/AI technology may create costs, such as additional hardware, for example, buying a cluster for handling Big Data or purchase of a fast computer with GPU support for training AI models. In addition, organizational structures may need to be changed or adapted. For example, if telemedicine data are automatically monitored, it requires a team of medical experts to check and decide if further actions need to be taken in case the system triggers an alarm.

Already in this phase, an ethical and legal governance model must be implemented that should be able to:

- Ensure that the AI system respects the values of human rights, human centric-approach, and human oversight, explainability, and fairness
- Ensure transparency at two levels:
- Internal:
    - Clearly identifying and notifying the roles and responsibilities of users, particularly concerning those uses involving the adaptation of decision-making processes subject to the risk of bias
    - Ensuring the involvement of users in continuous improvement
- External:
    - Providing adequate information to patients
    - Designing dialog methodologies with all relevant stakeholders
- Regularly audit the system from a legal and security point of view
- Maintain adequate incident management procedures, particularly those relating to security breaches and those identifying reliability issues in AI results.

Governance models may involve, depending on the characteristics of the entity:

- Adopting and implementing ethical codes
- Promoting and adopting the codes of conduct and/or certifications provided in the GDPR
- Defining governance bodies for the systems

In this phase, how information is presented and reported is also important. This also includes the development of a Graphical User Interface (GUI). Additionally, planning the deployment should include a fall-back strategy in case the new technology is not working as expected and interferes with the day-to-day operation.

This applies, in particular, to disruptive technologies, such as asset management technologies that replace older techniques.

The deployment of the developed Big Data/AI technology (the roll-out) should happen after the model at the IT infrastructure using the model and attached components, such as a newly developed GUI, has been thoroughly tested and evaluated. In contrast to most companies at which a roll-out of a new system can take place after a (partial-) shutdown of some services; in hospitals, usually, the IT system has to be running all the time and must only be interrupted for a short period and therefore should not be interfered by erroneous systems. If a backup twin system is available, it would be a good idea to test the system there first.

The team using the new system should be informed about the roll-out timely. Also, the outcomes of the model need to be communicated with the decision-makers, which were mainly included in the business understanding phase.

Monitoring is necessary to see how the newly integrated Big Data/AI technology behaves over time. Regarding Big Data, an aspect of monitoring is how fast the assigned storage is filled. Regarding AI models, an aspect of monitoring can be how the prediction or classification based on the new model behaves over time with new and, thereby, unseen data. For example, in a hospital, these data are usually based on patients (vitals, lab values, and so on), so monitoring may be necessary over a longer period because only a few new patients are being hospitalized every day. In addition, the deployment infrastructure should support the automatic collection of key KPIs that detect the performance of the solution being rolled out. This would allow AI models to be adapted after rollout and even detect any drifts that might impact the models' performance to make predictions. This also implies that it is critical to think about what KPIs need to be measured to accurately impact the system's rollout. This might also result in the need to integrate with other IT systems, which needs to be considered before deployment.

Maintenance may mean that a trained AI model is outdated and needs to be replaced either because the algorithm calculating the model has been significantly improved or due to a larger sum of new data collected, a new training, testing, and deployment of a model make sense. With regard to Big Data, the hardware infrastructure may be evaluated, and depending on the result, a Big Data cluster needs more computing power.

Reference

[1] Shearer C. The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing. 2000. 5(4), 13–22.