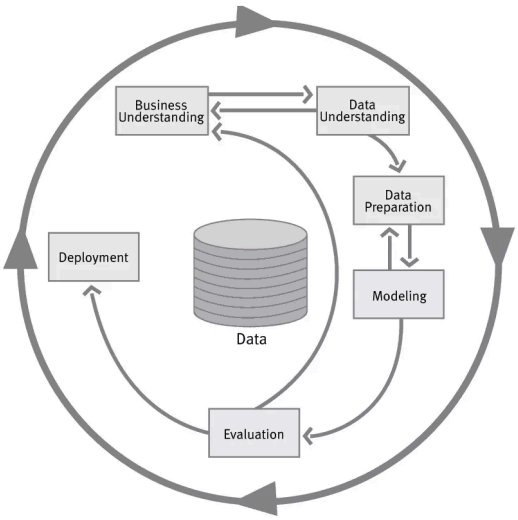
(https://datasolut.com)

CRISP-DM: Grundlagen, Ziele und die 6 Phasen des Data Mining Prozess



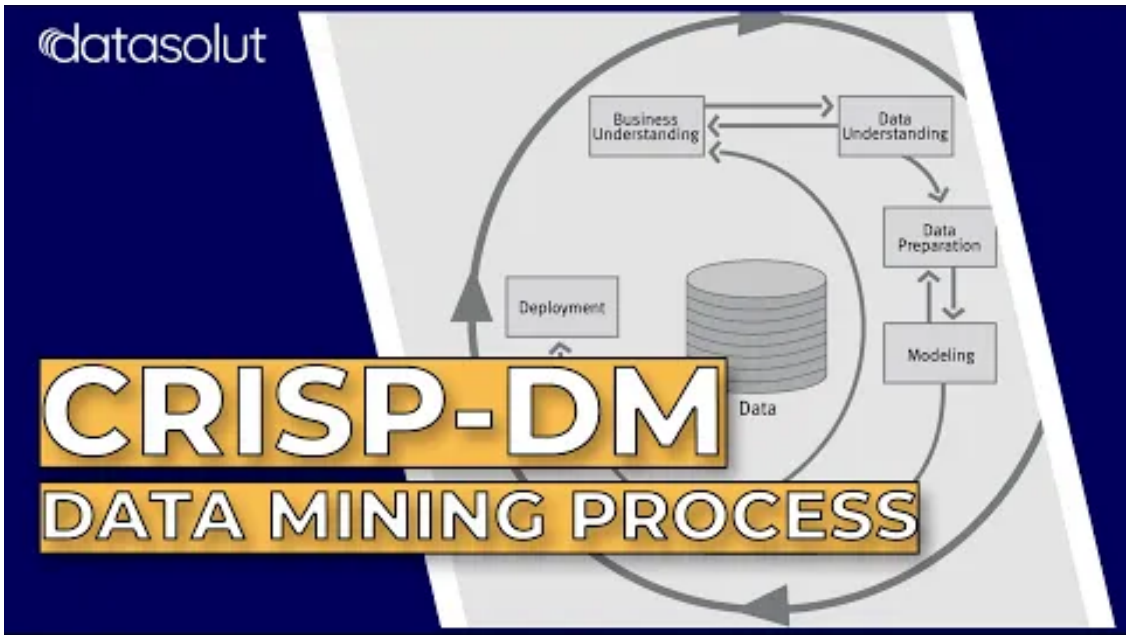

Laurenz Wuttke(https://datasolut.com/author/yxgy4p/)

Customer Analytics (https://datasolut.com/category/customer-analytics/), Machine Learning (https://datasolut.com/category/machine-learning/)

CRISP-DM ist ein einheitlicher Standard für die Entwicklung von Data Mining Prozessen und hilft Unternehmen dabei Data Mining Projekte gut zu strukturieren.

Haben Sie die Herausforderung ein Data Mining Projekt zu strukturieren? Oder wollen wissen was die „*best practices*“ für einen einheitlichen Prozess sind?

Für die Strukturierung von Data Mining oder Machine Learning Projekten nutze ich schon seit Jahren den **CRISP-DM Standard**. Dieser hilft vor allem dabei in Zusammenarbeit mit unseren Kunden den Data Mining Prozess genauer zu erklären und diesen in Ihre Unternehmensprozesse einzubinden.



Sie können sich CRISP-DM Modell auch in unserem Video dazu angucken.

In den folgenden Abschnitten gehe ich auf die Grundlagen des CRISP-DM ein und beschreibe die Phasen im Detail:

1. Was ist CRISP-DM?
2. Ziele von CRISP-DM
3. 6 Phasen eines Data Mining Projekts

1. Was ist CRISP-DM?

Im Jahr 2000 verschaffte man mit dem CRISP-DM Modell einen einheitlichen Standard für Data Mining Prozesse. Die Entstehung des CRISP-DM-Modells geht auf drei Unternehmen zurück, die sich seit dem Jahr 1996 als führende Organisationen in der Industrie der Auswertung großer Datenbestände, speziell dem Bereich Data Mining, gewidmet haben. Auch die EU war mit Fördermitteln an der Entwicklung beteiligt. (https://cordis.europa.eu/project/id/25959)

Das ursprüngliche CRISP-DM Paper finden Sie hier (https://www.the-modeling-agency.com/crisp-dm.pdf).

Diese haben durch kontinuierliche Weiterentwicklungen im Bereich Data Mining einen einheitlichen Standard mit dem CRISP-DM-Modell entwickelt. Gemeint sind damit die Unternehmen NCR System Engineering, SPSS Inc., und die DaimlerChrysler AG.

Die grundsätzliche Zielsetzung des CRISP-DM-Modells ist es, einen branchen-, software- und anwendungsunabhängigen standardisierten Prozessablauf des Data Minings (https://datasolut.com/was-ist-data-mining/)für Unternehmen bereitzustellen.

Ziele von CRISP-DM

Die Ziele von dem Data Mining Vorgehensmodell kurz zusammengefasst:

- Schaffung eines einheitlichen Prozess- und Vorgehensmodells für Data Mining Projekte
- Übergreifende Nutzung in verschiedenen Branchen
- Anleitung und Blaupause für Data Mining in 6 Schritten

6 Phasen eines Data Mining Projekts

Ergebnisse des Data Minings sollen durch das CRISP-DM-Modell schneller und präziser zur Verfügung stehen. Im Folgenden wird der CRISP-DM dargestellt, dieser ist in **sechs Schritte** unterteilt:

1. Business Understanding (Aufgabendefinition)
2. Data Understanding (Auswahl der relevanten Datenbestände)
3. Data Preparation (Datenaufbereitung)
4. Modeling (Auswahl und Anwendung von Data Mining Methoden)
5. Evaluation (Bewertung und Interpretation der Ereignisse)
6. Deployment (Anwendung der Ergebnisse)

Die einzelnen Phasen, sowie die Iterationen der einzelnen Phasen dieses Modells, lassen sich je nach Problemstellung unterschiedlich gewichten. Jede Phase dieses Modells spielt eine entscheidende Rolle für den Erfolg eines Data Mining-Projektes. In Abbildung 5 wird erkenntlich, dass das CRISP-DM-Modell einen Kreislauf darstellt und somit iterativ ist. Dies bedeutet, dass es keinen definierten Endpunkt in dem Data Mining-Prozess gibt. Der äußere Kreis stellt die Iteration des ganzen CRISP-DM-Modells dar, was bedeutet, dass jeder Durchlauf neue Fragen aufwerfen kann. Dabei trägt jede Wiederholung zu einer weiteren Optimierung des Prozesses bei.

CRISP-DM Modell nach Shear

Die inneren Pfeile machen deutlich, dass es sich bei dem Ablauf des Prozesses nicht um eine starre Sequenz handelt, sondern, dass es durchaus Rückkopplungen innerhalb des Zykluses geben kann. Dies kann passieren, wenn unvorhergesehene Probleme auftreten oder man Zwischenziele nicht in der gewünschten Qualität erreicht.

Phase 1: Business Understanding

Die erste Phase konzentriert sich auf die präzise Beschreibung der betriebs-wirtschaftlichen Problemstellung. Im nächsten Schritt eines Data Mining-Projektes überführt man die betriebswirtschaftlichen Problemstellungen in konkrete Anforderungen an die Datenanalyse. Diese bilden die zentrale Grundlage für alle weiteren Schritte und Entscheidungen im Data Mining-Prozess, z.B. über die Auswahl der Methoden.

Anhand der konkreten betriebswirtschaftlichen und analytischen Zielsetzungen ist ein Projektplan für das Projekt zu entwickeln. In dem Projektplan sind die erforderlichen zeitlichen, personellen und sachlichen Ressourcen zu spezifizieren.

Hierbei ist es wichtig, den Anwender mit in den Data Mining-Prozess einzubeziehen, um ein Verständnis der betriebswirtschaftlichen Fragestellung des Projektes zu entwickeln:

- **Bestimmung der betriebswirtschaftlichen Problemstellung:** Hier wird an den Anwender die Anforderung gestellt, das Data Mining Projekt betriebswirtschaftlich auszurichten. Dabei werden die operationalen und betriebswirtschaftlichen Zielkriterien formuliert.

- **Situationsbewertung:** Durch die Situationsbewertung werden vorhandene Software- sowie Personalressourcen, die für das Data Mining-Projekt zur Verfügung stehen, bestimmt. Zudem sind mögliche Risiken, die während des Data Mining-Projektes auftreten können, aufzuführen.
- **Bestimmung analytischer Ziele:** Ausgehend von der zuvor bestimmten betriebswirtschaftlichen Problemdefinition (z.B. zielgerichtete Ansprache der Kunden) müssen dazu die erforderlichen Datenanalyseaufgaben (z.B. Kundensegmentierung, Scoring-Verfahren zur Kampagnenoptimierung etc.) ermittelt werden. Zudem müssen die Erfolgskriterien für das Data Mining-Projekt bestimmt werden (z.B. Steigerung der Responsequote von Kampagnen um 3% bei weniger Ressourceneinsatz, Ansprache des Kunden nach Verhaltensmustern etc.).
- **Erstellung des Projektplans:** Der Projektplan beschreibt die beabsichtigten Ziele des Data Mining-Projektes, dazu gehört:
  - Auflistung der einzelnen Schritte mit Zeitspanne
  - Beurteilung möglicher Risiken (Verzögerungen, Ursachen für ein Scheitern des Projektes etc.)
  - Prüfung der zur Verfügung stehenden Ressourcen (wie Mitarbeiter ; Hardware; Software z.B. Data Warehouse, Data Mining-Werkzeuge; Datenbestand)

In Data Mining-Projekten wird in der Regel 50%-70% der Zeit für die Datenaufbereitung benötigt. Nur 20%-30% der veranschlagten Zeit entfallen dabei auf die Bestimmung der relevanten Datenbestände. Für die Modellierung, die Bestimmung der betriebswirtschaftlichen Fragestellung und Erfolgsmessung wird jeweils 10%-20% benötigt, lediglich 5%-10% der Zeit entfallen auf die Implementierung der erstellten Modelle.

Phase 2: Data Understanding

Nach der Formulierung der analytischen Ziele für das Data Mining ist nun eine Auswahl der relevanten Datenbestände zu treffen. Diese Phase dient dem Analysten, bestehende Zusammenhänge aus den Daten zu erkennen, eventuelle Qualitätsmängel der Daten festzustellen oder interessante Teilmengen zu identifizieren, um

eine Hypothese über die Daten aufzustellen. Die Phase besteht aus folgenden vier Schritten:

- **Daten sammeln:** Hier werden die benötigten Daten für die Analyse beschaffen und, wenn erforderlich, in bereits bestehende Datenmengen integriert. Der Analyst sollte Probleme, die bei der Datenbeschaffung auftreten, stets dokumentieren, um mögliche Diskrepanzen bei einem Folgeprojekt in der Zukunft zu vermeiden.
- **Daten beschreiben:** In diesem Schritt gilt es ein allgemeines Verständnis für die Daten zu erlangen. Zudem werden die Eigenschaften der Daten beschrieben, wie z.B. Quantität der Daten, Formateigenschaften, Anzahl der Einträge und Felder sowie Eigenschaften der Felder. Die entscheidende Frage die sich der Analyst stellen sollte ist, ob die vorliegenden Daten der Datenanalyse genügen um das Projekt erfolgreich abzuschließen.
- **Untersuchung der Daten:** Zur Untersuchung der Daten werden erste Analysen mit den Daten betrieben um z.B. bestimmte Produktgruppen zu identifizieren, die einen großen Teil des Umsatzes ausmachen. Hierzu werden Reports erstellt, um die ersten Erkenntnisse und Hypothesen zu visualisieren.
- **Bewertung der Daten:** An diesem Punkt wird die Qualität des Datenbestandes bewertet. Es sollte festgestellt werden, ob die Datenmenge für die Analyse ausreichend und verwendbar ist. Besonders ist auf fehlende Attributwerte zu achten.

Phase 3: Data Preparation

Die Data Preparation-Phase umfasst alle Aktivitäten zur Erstellung der finalen Datenmenge oder Datenauswahl, die zur Analyse in die Modellierungssoftware geladen wird. Hierbei konzentriert sich der Anwender auf die Auswahl der Tabellen, Einträge und der Attribute sowie insbesondere auf die Transformation und Bereinigung der Daten. Im Folgenden werden die Schritte der Datenaufbereitung beschrieben:

- **Auswahl der Daten:** Die Auswahl der Daten für das Data Mining hängt stark von den Zielen ab, die man für das Data Mining-Projekt definiert. Hier spielen die Datenqualität und die technischen Gegebenheiten eine große Rolle. Es wird eine Selektion der Daten vorgenommen, wie z.B. eine Auswahl aller Kunden, die einen Umsatz von mehr als 100 Euro im Monat generieren. Am Ende dieses Prozesses sollte sich deutlich zeigen, welche Datenmengen(-Sets) in die Analyse aufgenommen werden oder ausgeschlossen werden.
- **Bereinigung der Daten:** Ohne eine Bereinigung der Daten ist ein erfolgreiches Data Mining-Projekt fraglich. Es gilt eine saubere Datenmenge auszuwählen oder die Datenmenge muss bereinigt sein, um das gewünschte Ergebnis in der Modellierung zu erreichen.
- **Transformation und Integration der Daten:** Um die Daten in eine brauchbare Darstellungsform zu bringen, transformiert man die Daten. Die Transformation kodiert Daten und verändert deren Granularität durch Aggregation oder Disaggregation. Wichtige Kennzahlen, die für eine Analyse zu erstellen sind, könnten z.B. Umsatz pro Kunde, Deckungsbeitrag pro Kunde oder Umsatzanteil in Produktgruppe pro Kunde etc. sein.
- **Format Data:** In einigen Fällen muss für die Modellierung eine einfache Anpassung des Datenformates erfolgen, z.B. Anpassung des Datentyps.

Phase 4: Modeling

Für eine betriebswirtschaftliche Problemstellung können in der Regel mehrere Modellierungstechniken des Data Mining zum Einsatz kommen. Einige Techniken stellen besondere Anforderungen an die Datenstruktur. Dies kann zur Folge haben, dass man von der Modellierung einen Schritt zurück in die Phase Data Preparation gehen muss. Dort geschieht ggf. eine Anpassung an Format oder Struktur der Daten:

- **Auswahl der Modellierungstechnik:** Hier gilt es eine Modellierungstechnik auswählen, mit der man das Modell erstellt
- **Testmodell erstellen:** Nach Auswahl des Modells wird ein Testmodell erstellt, um Qualität und Genauigkeit des Modells zu überprüfen. In überwachten Verfahren, wie der Klassifikation, ist es üblich, die Fehlerraten als Qualitätsmaß zu nutzen.
- **Bewertung des Modells:** Hier ist das Modell nach der im Vorfeld definierten Data Mining-Zielsetzung zu bewerten. Des Weiteren gilt es die Data Mining-Ergebnisse in Bezug auf die betriebswirtschaftliche Fragestellung zu bewerten.

Phase 5: Evaluation

Bevor das Modell zum Einsatz kommt, ist es wichtig das Modell zu bewerten. Es ist zu hinterfragen, ob das Modell wirklich die Qualität bietet, um der Zielsetzung des Data Mining-Projektes zu genügen. Es ist zu bewerten, ob das Modell wirklich der Zielsetzung des Data Mining-Projektes genügt. Lassen sich die Ziele nicht erreichen, findet gegebenenfalls ein erneuter Durchlauf der Phase statt. Folgende Schritte sind Aufgaben der Evaluationsphase:

- **Bewerten der Resultate:** In diesem Schritt bewertet man, inwieweit das Modell die Projektziele erreicht. Wenn die Ziele nicht erreicht sind, ist aufzuführen, aus welchen Gründen.
- **Bewertung des Prozesses:** Das Data Mining-Projekt wird rückblickend bewertet. Es wird festgestellt, ob alle wichtigen Faktoren betrachtet wurden und inwieweit die Attribute für zukünftige Data Mining-Projekte zu nutzen sind.
- **Nächste Schritte festlegen:** In diesem Schritt entscheidet der Projektleiter, ob das Projekt beendet ist und eingeführt wird.

Phase 6: Deployment

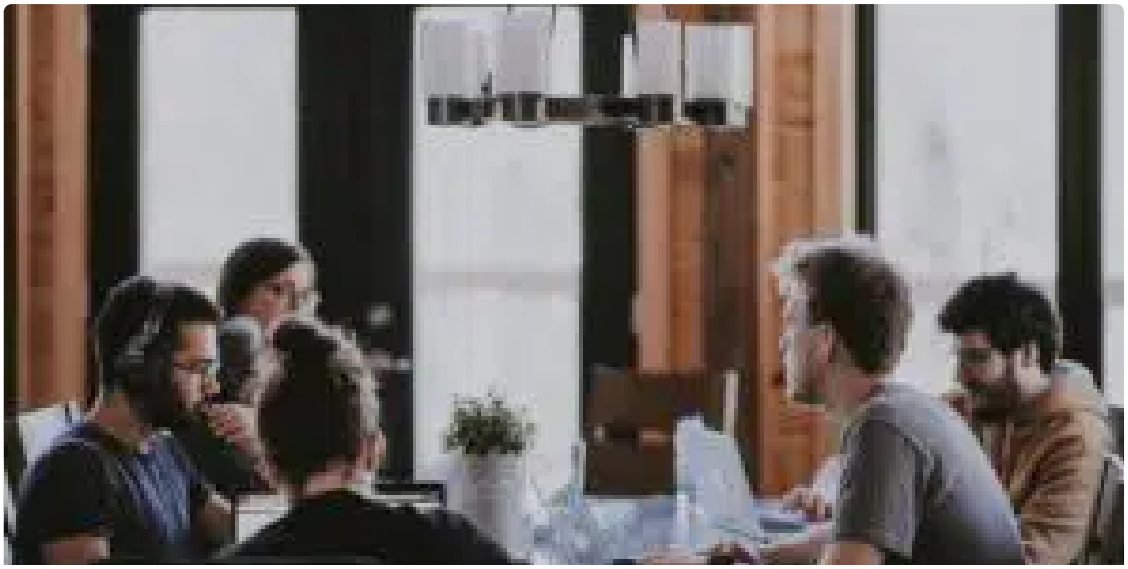
Die Deployment-Phase bildet in der Regel die Endphase eines Data Mining-Projektes. Hier werden die gewonnenen Erkenntnisse so geordnet und präsentiert, sodass für den Auftraggeber die Möglichkeit besteht dieses Wissen zu nutzen. Dazu gehört eine eventuelle Implementierungsstrategie, die Überwachung der Gültigkeit der Modelle, ein zusammenfassender Bericht und eine Präsentation.

Auch interessant für Sie

(https://datasolut.com/kundenwert/)

Kundenwert: wie wertvoll ist jeder einzelne Kunde? (https://datasolut.com/kundenwert/)

Weiterlesen » (https://datasolut.com/kundenwert/)



(https://datasolut.com/aufbau-eines-data-science-teams/)

Aufbau eines Data Science Teams (https://datasolut.com/aufbau-eines-data-science-teams/)

Weiterlesen » (https://datasolut.com/aufbau-eines-data-science-teams/)

(https://datasolut.com/marketinganalysen/)

Marketinganalysen: Vorteile, Methoden und Herausforderungen (https://datasolut.com/marketinganalysen/)

Weiterlesen » (https://datasolut.com/marketinganalysen/)

(https://datasolut.com)

datasolut GmbH - Mehr Wert mit KI.

hello@datasolut.com

+49 221 7040365

(https://www.linkedin.com/company/datasolut-gmbh) (https://www.youtube.com/channel/UCsztRKWmTuScsxFy0y59l0w/) (https://www.xing.com/companies/datasolutgmbh) (https://www.facebook.com/datasolut.offical/)

Mitglied im KI-Bundesverband:

(https://ki-verband.de/)

Offizieller Technologiepartner:

databricks Partner

(https://www.databricks.com/company/partners/consulting-and-si)

Navigation

Lösungen(https://datasolut.com/loesungen/)

KI Use Cases(https://datasolut.com/ki-use-cases/)

Technologiepartner(https://datasolut.com/Technologiepartner/)

Über uns(https://datasolut.com/ueber-uns/)

Karriere(https://datasolut.com/karriere/)

Kontakt(https://datasolut.com/kontakt/)

Blog(https://datasolut.com/blog)

datasolut Wiki(https://datasolut.com/wiki)

Unsere Lösungen

Data Science Beratung(https://datasolut.com/loesungen/data-science-beratung/)

Data Engineering Beratung(https://datasolut.com/loesungen/data-engineering-beratung/)

Churn Management(https://datasolut.com/loesungen/churn-management/)

Customer Lifetime Value (CLV)(https://datasolut.com/loesungen/customer-lifetime-value/)

Next Best Offer (NBO)(https://datasolut.com/loesungen/next-best-offer/)

Forecasting(https://datasolut.com/forecasting/)