Duale Hochschule Baden-Württemberg
Mannheim

Bachelorarbeit

Modellierung einer Funktion zur Berechnung der Wahrscheinlichkeit eines Torerfolges im Fußball

Studiengang Wirtschaftsinformatik

Studienrichtung Software Engineering

Verfasser:	Alexander Baum
Matrikelnummer:	8095497
Firma:	SAP SE
Abteilung:	SAP Sports
Kurs:	WWI 14 SE A
Studiengangsleiter:	Prof. Dr.-Ing. Jörg Baumgart
Wissenschaftliche Betreuerin:	Susanne Klusmann susanne.klusmann@f-i.de +49 511 5102-22137
Firmenbetreuer:	Dr. Andrew McCormick-Smith andrew.mccorcmick-smith@sap.com +49 6227 7-41565
Bearbeitungszeitraum:	21. November 2016 bis 20. Februar 2017

Kurzfassung

Verfasser: Alexander Baum

Kurs: WWI 14 SE A

Firma: SAP SE

Thema: Modellierung einer Funktion zur Berechnung der Wahrscheinlichkeit eines Torerfolges im Fußball

- Problemstellung - Ziele - Vorgehen - Ergebnisse

Inhaltsverzeichnis

Verzeichnisse

Abkürzungsverzeichnis

CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
IoT	Internet of Things
KDD	Knowledge Discovery in Data
OLAP	Online Analytical Processing
SAP	eigenständiger Markennamen - früher: *Systeme, Anwendungen und Produkte in der Datenverarbeitung*
SQL	Structured Query Language

Abbildungsverzeichnis

Tabellenverzeichnis

Listingverzeichnis

1 Einleitung

1.1 Ziel

Hintergrund: - Begriff Expected Goals wird als einer der neuen Schlüsselindikatoren im Fußball angesehen - Frage nach der Wahrscheinlichkeit von Punkt X,Y einen Torerfolg zu erzielen - Zugrunde liegen die Spieldaten der Bundesligasaisons 2014/15, 2015/16, sowie die aktuellen Spiele der Saison 2016/16 - Expected Goals gibt es in zahlreichen Varianten, doch wurde noch keine Funktion dafür modelliert (Ziel der Arbeit = neues Wissen schaffen) - Trainer, Spielanalysten und Scouts würden von einem fundierten und wissenschaftlich begründeten KPI profitieren

- Beantwortung der Fragen: 1. Welche Daten liegen vor? 2. Wie sollen die für die Funktion relevanten Daten selektiert werden? 3. Müssen Daten bereinigt bzw. aufbereitet werden? 4. Wie kann eine Funktion aus Daten modelliert werden? 5. Welche Arten der Regressionsanalyse gibt es? 6. Welche Tools/welche Software kann für die Berechnung genutzt werden? 7. Welche Annahmen werden für das Modell getroffen und warum? 8. Wie kann der Erfolg der resultierenden Funktion gemessen werden?

1.2 Umgebung

Unternehmen Die SAP[1] wurde 1972 von fünf ehemaligen IBM Mitarbeitern gegründet und ist seit mehr als 40 Jahren, hinsichtlich des Marktanteils mit über 282.000 Kunden, das weltweit führende Unternehmen für Anwendung- und Analysesoftware. Der im baden-württembergischen Walldorf gegründete Aktienkonzern bietet mit dem bis heute bekanntesten Produkt *SAP ERP* eine Softwarelösung zur Abbildung aller Geschäfts- und Produktionsprozesse in einem Unternehmen von

[1] *eigenständiger Markennamen - früher: Systeme, Anwendungen und Produkte in der Datenverarbeitung* (*SAP*)

Personal- und Rechnungswesen bis hin zur Logistik. Mit dem heutigen Stand der Entwicklung setzt die SAP ihren Fokus verstärkt auf die Bereiche Cloud, Mobile und Internet of Things, um mit den anderen Unternehmen konkurrieren zu können und den Anschluss an den Trend der Zeit nicht zu verlieren. Die SAP beschäftigt in über 180 Ländern mehr als 77.00 Mitarbeiter und erzielte im Jahr 2015 einen Umsatz von 20,8 Mrd Milliarden Euro, sowie ein Betriebsergebnis von 6,3 Milliarden Euro.[2]

Abteilung Die Praxisphase erfolgte in der Abteilung *Sports & Enternainment*, die sich von den klassischen SAP Geschäftsbereichen isoliert hat und alles rund um den Sport betreut. Im Bereich des Fußballs liegt der Fokus einerseits auf der Organisation des gesamten Vereins inklusive Umfeld, sprich Management, Marketing, Mannschaft, Jugend oder auch Fans, andererseits auch auf der Spielanalyse mit Hilfe von erhobenen Daten. Dazu steht die Abteilung in regelmäßigen Kontakt mit dem Bundesligaverein der TSG 1809 Hoffenheim sowie der deutschen Nationalmannschaft, um ständig neue Anwendungsfälle zu gewinnen. Alle Funktionalitäten sollen in einem Produkt, dem sogenannten *Sports One* vereint werden, welches aus verschiedenen Rollen, wie Spieler, Trainer oder auch Mannschaftsarzt verwendet werden kann. Im Bereich der Spielanalyse und der Leistungsdiagnostik werden Unmengen an Daten gesammelt, die es für den späteren Anwender zu visualisieren gilt. Hier findet sich der in dieser Arbeit beschriebene Anwendungsfall wider, mit dessen unterstützender Funktion eine Funktion für die Berechnung der Wahrscheinlichkeit eines Torerfolges modelliert werden soll.

1.3 Vorgehen

Methodik: Als grundlegende Methodik wird der allgemeingültige Knowledge Discovery Process verwendet. Der Fokus liegt dabei vor allem im Schritt des Data Minings, in dem auch die Funktion letztendlich modelliert wird. Die vorherigen Schritte zeigen die Datenaufbereitung als auch die –transformation, um den ganzen Kontext besser verstehen zu können. In den einzelnen Schritten gibt es wiederum wissenschaftliche Methoden, die im theoretischen Teil kurz vorgestellt und in der Umsetzung dann angewendet werden. Beispielsweise findet sich unter dem Punkt Data Mining die mathematische Methode der Regressionsanalyse. So kann der Leser die Arbeit systematisch nachvollziehen und sich entlang des roten Pfadens hangeln.

[2] Zahlen vor Abzug der Steuern
Weitere Information zum Geschäftsbericht der SAP SE aus dem Jahr 2015 unter:
http://www.sap.com/docs/download/investors/2015/sap-2015-geschaeftsbericht.pdf
[10.01.2017]

Erwartete Ergebnisse: - verschiedene Funktionen bei unterschiedlicher Betrachtung: o der Auswahl der Daten (Schüsse aus dem Spiel, Standards, ...) o des Winkels zum Tor o der Distanz zum Tor - unterschiedliche Flächen der Funktion im dreidimensionalen Raum o Kegel o Teil eines Ellipsoids

2 Theoretische Grundlagen

2.1 Data Mining

Die vorliegende wissenschaftliche Fragestellung bewegt sich im Bereich des Data Minings. Das folgenden Kapitel soll dem Leser dazu eine Einführung in die Thematik geben, um ein grundlegendes Verständnis der Begriffe und Ziele des Data Minings zu erlangen Kapitel 2.1.1. Darüber hinaus werden die Prozesse des Data Minings Kapitel 2.1.2 auf S. 7 beleuchtet, wobei der *Knowledge Discovery in Data* Prozess – methodischer Aufbau der späteren Umsetzung – in Kapitel 2.2 auf S. 11 nochmal ausführlich eruiert wird.

2.1.1 Definition des Data Minings

Der Bergriff des Data Minings reicht zurück bis in die 80er Jahre des letzten Jahrhunderts und verfolgt das Ziel, Wissen aus riesigen Datenmengen zu extrahieren.[3] Als ein Prozess des *Sammelns, Säuberns, Verarbeitens und Analysierens von Daten, zur Gewinnung von nützlichen Informationen*[4] beschreibt Aggarwal diesen Begriff. Denn der immense Datenanstieg in den letzten Jahrzehnten erlaubt uns nicht einfach wertvolle Informationen oder organisiertes Wissen automatisch zu verstehen oder zu entnehmen. Das heutige „Informationszeitalter“ führte zum Beginn des renommierten Wissenschaftsbereich des Data Minings, der in der Literatur auch als natürliche Evolution der Informationstechnologie bezeichnet wird.[5,6] Grundlegende interdisziplinäre, wissenschaftliche Teilgebiete des Data Minings sind, z.B. Statistik, maschinelles Lernen, Mustererkennung, Systemtheorie oder künstliche Intelligenz.[7,8]

[3] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen intelligenter Datenanalyse, 2015, S.2.
[4] Vgl. *Aggarwal*, Data mining: The textbook, 2015, S.1.
[5] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.1.
[6] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.2.
[7] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen intelligenter Datenanalyse, 2015, S.2.
[8] Vgl. *Shi* et al., Intelligent knowledge: A study beyond data mining, 2015, S.1.

Cleve und Han vergleichen die Suche nach Muster und Zusammenhänge in den Daten mit dem Abbau von Rohstoffen.[9] Sowie im Bergbau nach Schätzen wie Gold und Silber im Gestein sucht wird, so strebt das *Data Mining* (*DM*) nach dem Ableiten von Wissen aus den (Roh-)Daten.[10,11] Ha9n geht sogar einen Schritt weiter und präferiert den Begriff des *knowledge mining from data* – referenzierend auf den verwendenden Terminus des *gold mining* statt des *rock or sand mining* – da diese Bezeichnung das eigentliche Ziel der Gewinnung von Wissen beinhaltet.[12,13]

„Unter Wissen verstehen wir interessante Muster, die allgemein gültig sind, nicht trivial, neu, nützlich und verständlich.“[14] Insofern wird das Ziel verfolgt komplexe Paradigmen zu erkennen, die durch die blanke Betrachtung der Daten nicht aufgedeckt werden könnten. Oftmals fehlt dem Datenanalyst das spezifische Fachwissen zur Erkennung von Mustern, sodass durch die Einbeziehung von Experten ein iterativer Prozess entsteht, bis ein gewünschtes Ergebnis erzielt wird. Zunächst werden aus den Daten, Informationen gewonnen, aus welchen wiederum das Wissen abgeleitet werden kann, wobei in diesem Prozess der Wissensextraktion die Datenmenge sukzessive abnimmt und sich verdichtet, wie in Abbildung 1 verdeutlicht.

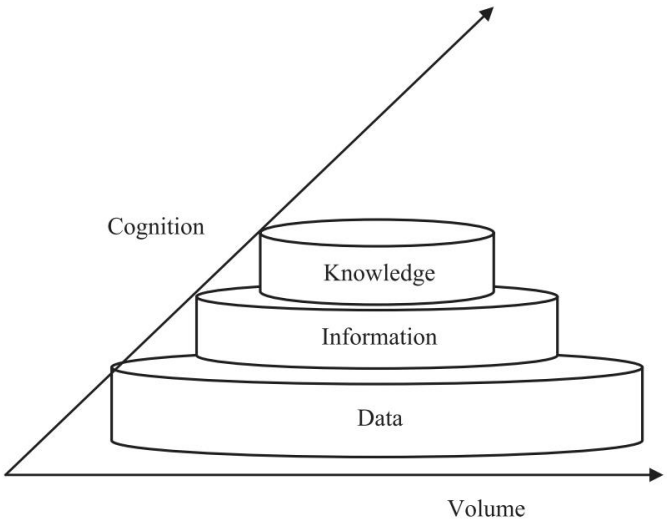

Abbildung 1: Wissensextraktion aus Daten[15]

[9] Die englische Übersetzung lautet *„Mining“*
[10] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.1.
[11] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.5-6.
[12] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.5-6.
[13] Weitere Termini nach Han: *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.*
[14] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen intelligenter Datenanalyse, 2015, S.2.
[15] Vgl. Abbildung *Shi* et al., Intelligent Knowledge, 2015, S.5

Durch den Einsatz von modernster Computerhard- als auch software ist es möglich immens große Datenmenge zu erheben, zu verarbeiten und zu analysieren, wodurch in diesem Kontext der Begriff *Big Data* entstanden ist.[16] Big Data bezeichnet Datenmengen, die mit herkömmlichen Analysemethoden nicht mehr zu verarbeiten werden und den Einsatz von Data Mining benötigen.[17,18] Dazu ein paar ausgewählte Beispiele aus verschiedenen Datenbereichsquellen:[19]

- **Word Wide Web:** Die Anzahl der Dokumente im Internet hat schon lange die Milliarden Marke geknackt, wobei die des unsichtbaren Webs noch viel größer ist. Durch Nutzerzugriffe auf Inhalte werden auf Serverseite Log-Dateien kreiert, um beispielsweise die Auslastung und Zugangszeiten zu protokollieren. Andererseits wird das Kundenverhalten auf kommerziellen Seiten aufgezeichnet, um personalisierte Werbung schalten zu können.

- **Benutzerinteraktion:** Festnetzanbieter nutzen die durch Telefonate entstanden Daten, wie Gesprächslänge und Ort, um relevante Muster über die Netzwerkauslastung, zielgerichtete Werbung oder auch anzusetzende Preise durch Datenanalyse zu extrahieren.

- **Internet of Things:** Durch kostengünstige (tragbare) Sensoren und deren kommunikative Vernetzung entstand das *Internet of Things* (*IoT*). Einer der Trends der heutigen Informationstechnologie, der durch die Erhebung von Massendaten eine signifikante Rolle für das Data Mining einnimmt.

- **Weitere Beispiele:** Social Media Plattformen (allen voran Facebook, Twitter und Co.), Finanzmärkte (z.B. der Aktienmarkt), Sport (z.B. Baseball, Basketball, Football oder wie in dieser Arbeit Fußball), uvm.[20,21]

Wir befinden uns in einer Situation, in der wir reich an Daten sind, jedoch arm an Informationen und Wissen. Der unglaubliche rasante und gigantische Datenzuwachs hat bei langem unsere menschliche Vorstellungskraft und Möglichkeiten übertroffen, sodass wir auf mächtige Werkzeuge angewiesen sind.(siehe Kapitel 2.2.4 auf S. 17) Die sich immer weiter ausbreitende Lücke zwischen Daten und Information führt nur durch die Nutzung von Methoden des Data Minings zu den „*Golden Nuggets of Knowledge*“.[22] Dazu müssen die (Roh-)Daten gezielt ausgewählt und umstrukturiert

[16] Vgl. *Witten/Frank/Hall*, Data mining: machine learning and techniques, 2011, S.3.
[17] Vgl. *Fasel/Meier*, Big Data: Grundlagen, Systeme und Nutzungspotenziale, 2016, S.5.
[18] Vgl. *Shi* et al., Intelligent knowledge: A study beyond data mining, 2015, S.1.
[19] Vgl. *Aggarwal*, Data mining: The textbook, 2015, S.2.
[20] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, AI Magazine 17 [1996], 1996, S.39.
[21] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.1-2.
[22] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.5.

werden, um diese anschließend von Algorithmen analysieren zu können. Folglich entstanden Data Mining Prozesse, die dieses Problem mit Hilfe systematischer Abläufe lösen sollen.(vgl. Kapitel 2.1.2) Zudem wird „Data Mining [...] heute durch eine zunehmende Anzahl von Software-Tools unterstützt, z. B. KNIME, MATLAB, SPSS, SAS, STATISTICA, TIBCO Spotfire, R, Rapid Miner, Tableau, QlikView, oder WEKA.“[23] Das Software-Tool *MatLab* wird innerhalb der Funktionsmodellierung in Kapitel 2.3 auf S. 17 vorgestellt und anschließend als Werkzeug zur Nutzung von Data Mining Methoden in der Umsetzungsphase genutzt.(vgl. Kapitel 4 auf S. 19)

2.1.2 Data Mining Prozesse

In der Literatur distinguieren viele Wissenschaftler den Begriff des eigentlichen Data Minings, gegenüber dem Gesamtprozess der Extraktion von Wissen. Andere wiederum behandeln beide Termini synonym zu einander.[24,25,26] Schlechte Qualität der Daten mindert die Leistungsfähigkeit des Data Minings. Um die Aussagekraft der Daten nicht zu gefährden, sind vorab Prozessschritte notwendig, die Daten in geeigneter Form für die Methoden des Data Minings bereitstellen.[27] Hierzu werden im folgenden kurz die zwei bekanntesten Prozessmodelle vorgestellt:

- *Knowledge Discovery in Data* (*KDD*)
- *Cross Industry Standard Process for Data Mining* (*CRISP-DM*)

2.1.2.1 Knowledge Discovery in Data

Der Begriff des *Knowledge Discovery in Data* Prozesses wurde in den frühen 90er Jahren geprägt und wird als „nicht trivialer Prozess zur Identifizierung von gültigen, neuartigen, potentiell sinnvolle und letztlich verständlichen Muster in Daten“[28,29] definiert.[30] Erstmals wurde der Terminus von Gregory Piatetsky-Shapiro auf der

[23] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen intelligenter Datenanalyse, 2015, S.3.
[24] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, AI Magazine 17 [1996], 1996, S.39.
[25] Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.2.
[26] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.1.
[27] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.10.
[28] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.1-2.
[29] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, AI Magazine 17 [1996], 1996, S.41.
[30] Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.2.

International Joint Conference on Artificial Intelligence, 1989 in Detroit (USA), formuliert und vorgestellt.[31] Der in Abbildung 2 abgebildete iterative KDD-Prozess nach Fayyad beinhaltet folgende Schritte, wobei das DM als ein eigener Prozessschritt ausgewiesen wird:[32]

- **Datenselektion**: Auswahl der geeigneten Datenmengen
- **Datenvorverarbeitung**: Behandlung fehlender oder problembehafteter Daten
- **Datentransformation**: Umwandlung in adäquate Datenformate
- **Data Mining**: Suche nach Muster
- **Interpretation und Evaluation**: Interpretation der Ergebnisse und Auswertung

Auf die einzelnen Prozessschritte und deren Methoden wird genauer in Kapitel 2.2 auf S. 11 eingegangen.

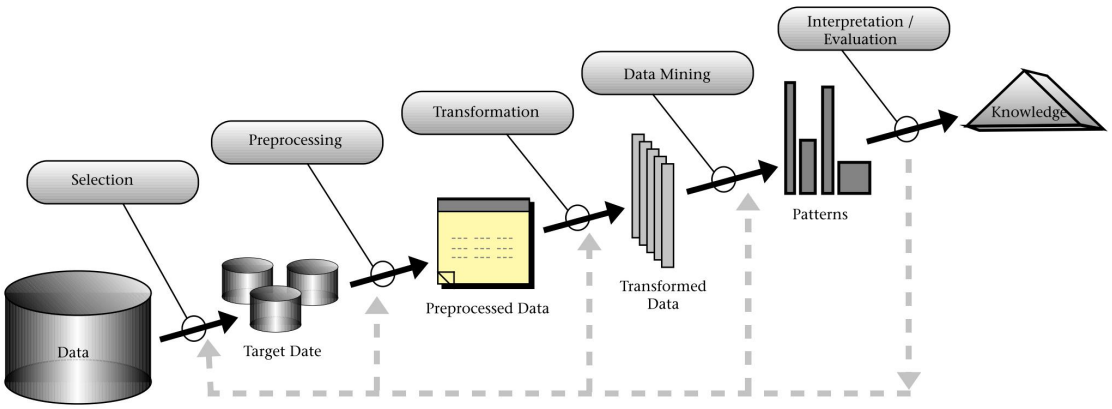

Abbildung 2: Der Knowledge Discovery in Data Prozess[33]

[31] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S.1.
[32] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.5.
[33] Vgl. Abbildung *Fayyad* et al., From Data Mining to Knowledge, 1996, S.41

2.1.2.2 CRISP-DM

Das CRISP-DM-Modell wurde im Jahr 200 durch ein Konsortium, bestehend aus mehreren Firmen, entwickelt. Beteiligt waren daran:[34,35]

- NRC Corporation,
- Daimler AG,
- SPSS,
- Teradata und
- OHRA.

Dieses Modell verfolgt das Ziel, einen standardisierten und branchenübergreifenden Data Mining Prozess zu definieren und das dadurch berechnete Modell zu validieren. Hierbei wird von einem Lebenszyklus mit den folgenden sechs Etappen ausgegangen, die in Abbildung 3 dargestellt werden.[36]

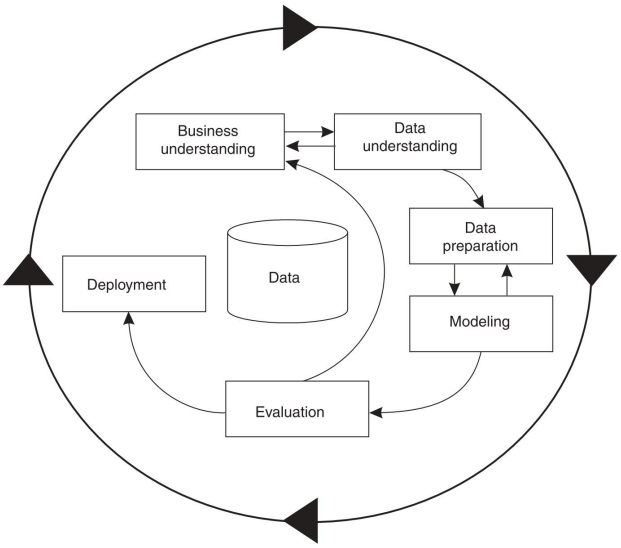

Abbildung 3: CRISP-DM Prozess[37]

[34]Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.6.
[35]Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.3.
[36]Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.6-8.
[37]Vgl. Abbildung *Mariscal* et al., A survey of data mining, 2010, S.13

Verstehen der Aufgabe Hier steht das grundsätzliche Verständnis des Fachgebiets und der Aufgabe im Vordergrund. Die Ziele werden abgestimmt, Ressourcen des Unternehmens ermittelt und die Ausgangssituation bestimmt. Weiterhin müssen Erfolgskriterien und Risiken quantifiziert werden, um eine Kostenplanung aufstellen zu können.

Verständnis der Daten Die Phase beschäftigt sich mit den benötigten Daten für das Ziel der Analyse. Daten werden gesammelt und beschrieben, um deren betriebliche Bedeutung zu verstehen.

Datenvorbereitung Es gilt den Data Mining Schritt vorzubereiten, wobei fehlerhafte und inkonsistente Daten korrigiert werden müssen, um diese schließlich in eine Datenstruktur zu transformieren, die für die Methoden des Data Minings nutzbar sind.

Data Mining - Modellbildung In dieser Phase wird ein Modell mit Hilfe des Data Minings erstellt, welches durch ein iteratives Verfahren immer wieder verfeinert und verbessert wird.

Evalution Die erzielten Ergebnisse werden mit den aus Phase 1 erstellten Erfolgskriterien gemessen, um beispielsweise festzustellen, ob der wirtschaftliche Nutze erzielt wurde.

Einsatz im Unternehmen Zuletzt wird der Einsatz der Resultate in das Unternehmen vorzubereiten und in das operative Geschäft zu integrieren.

Das Modell bezieht und orientiert sich, wie schon im Namen zu lesen, stark an wirtschaftlichen Projekten und beschreibt *was* zu tun ist, aber nicht genau *wie*, sodass Projektteams innerhalb dieses Rahmens beginnen ihre eigenen Methoden zu verwenden.[38]

Im Vergleich zum KDD-Modell nach Fayyad, sind die Phase 1 und 2 des CRISP-DM-Modells sehr stark projektabhängig und spiegeln die Sicht der Industrie auf das Projekt wider.[39] Im Gegensatz dazu konzentriert sich der KDD-Prozess auf die

[38] Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.4.
[39] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.8.

Datenbereitstellung und Analyse, sodass dieser als grundlegende Methodik für die spätere Umsetzung der wissenschaftlichen Aufgabenstellung herangezogen wird und genauer in Kapitel 2.2 durchleuchtet wird.

Mariscal et al. diskutieren in ihrer Studie weitere zahlreiche Prozessmodelle zur Extraktion von Wissen aus riesigen Datenmengen, wobei die Kernelemente der Datenselektion, -vorverarbeitung und -tranformation, sowie der anschließende Schritt des eigentlichen Data Minings immer wieder aufzufinden sind.[40] Nicht zuletzt sei zu erwähnen, dass in der Literatur unterschiedliche Auffassungen zu dem Begriff *Data Mining* existieren und dieser oftmals mit den Data Mining Prozessen synonym verwendet wird. Ein Hinweis darauf sind auch die weit über 500 wissenschaftliche Artikel zu dem Journal *Data Mining and Knowledge Discovery* auf Springer Link.

2.2 Knowledge Discovery in Data

Das folgende Kapitel beschreibt den *Knowledge Discovery in Data* Prozess, der im vorherigen Kapitel (vgl. Kapitel 2.1.2.1 auf S. 7) als grundlegende Methodik der Arbeit ausgewählt wurde. Hierzu werden die einzelnen Prozessschritte der Datenselektion, der Datenvorverarbeitung, der Datentransformation, der Data Mining Methoden, sowie der Interpretation der Ergebnisse näher durchleuchtet, um diese in der späteren Umsetzung der wissenschaftlichen Aufgabe nutzen zu können.

„Experten [...] haben realisiert, dass seine große Anzahl an Datenquellen der Schlüssel zu bedeutsamen Wissen sein kann und das dieses Wissen in dem Entscheidungsfindungsprozess genutzt werden sollten. Eine einfache *Structured Query Language* (*SQL*)-Abfrage oder *Online Analytical Processing* (*OLAP*) reichen für eine komplexe Datenanalyse oft nicht aus.“[41] Hier greift der in Abbildung 2 auf S. 8 dargestellte KDD-Prozess, ein multiples iteratives Modell, indem die einzelnen Schritte solange wiederholt und gegeneinander abgestimmt werden müssen, bis aus den zugrundeliegenden Daten, Wissen abgeleitet werden kann.[42] Das Data Mining selbst kommt erst nach ausführlicher Datenvorbereitung zum Einsatz und kann so zu einer automatischen und explorativen Anpassung eines Modells – wie der Funktionsmodellierung (vgl. Kapitel 2.3 auf S. 17) – an riesige Datenmengen genutzt werden.[43,44]

[40] Vgl. vorgestellte Modelle *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010.
[41] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S.1.
[42] Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.7.
[43] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S.1.
[44] Vgl. *Mariscal/Marbán/Fernández*, A survey of data mining processes, 2010, S.7.

In der Literatur existieren unterschiedliche Vorstellungen der einzelnen Prozessschritte, wodurch es oftmals zu Überschneidungen zwischen den einzelnen Gebieten gibt. So findet sich die Methode der *Data Integration* einerseits in der Datenselektierung wieder, andererseits auch in der Datenvorverarbeitung.[45,46] Im Folgenden wird versucht diese Schritte klar von einander abzutrennen.

2.2.1 Datenselektion

Die Datenselektion befasst sich hauptsächlich mit der Auswahl der geeigneten Datenmengen – der *Zieldaten* – auf denen die spätere Erforschung ausgeübt wird.[47] Der Datenanalyst befasst sich in dieser Phase mit der Bestimmung der für die Analyse geeigneten Daten und des Exports dieser Datenauswahl in beispielsweise einer Datenbank. Die selektierten Daten können beispielsweise technischen oder rechtlichen Restriktionen unterliegen, wie zum Beispiel Zugriffs- oder Kapazitätsbeschränkungen. Hierbei sollte auf eine repräsentative Teilmenge des Datenbestands zurückgegriffen werden.[48]

2.2.2 Datenvorverarbeitung

„Da die Zieldaten aus den Datenquellen lediglich extrahiert werden, ist im Rahmen der Datenvorverarbeitung die Qualität des Zieldatenbestands zu untersuchen und – sofern nötig – durch den Einsatz geeigneter Verfahren zu verbessern.“[49]

Diese essentielle Phase verfolgt das Ziel, die unstrukturierten und zunächst nutzlos scheinenden selektierten Rohdaten, in Daten höherer Qualität umzuwandeln, um diese der passenden DM-Methode im geeigneten Format bereitzustellen zu können. Die Struktur und das Format muss perfekt auf die vorliegende Aufgabe passen, ansonsten führt die geringe Qualität der Daten zu schlechten bzw. falschen Resultaten bis hin zu Laufzeitfehlern.[50] Es gilt auch hier das alte Prinzip: GIGO – garbage in, garbage out.[51] Die oftmals schlechte Qualität der (Roh-)Daten ist durch *feh-*

[45] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.1.
[46] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.198.
[47] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, AI Magazine 17 [1996], 1996, S.42.
[48] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.9.
[49] *Cleve/Lämmel*, Data Mining, 2014, S.9.
[50] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.10-11.
[51] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.197.

lende, ungenaue, inkonsistente bzw. widersprüchliche Daten zu begründen.[52,53] Im Folgenden werden dazu einige Ursachen beispielsweise aufgeführt.

Ungenaue bzw. falsche Daten können schon bei der Erhebung entstehen, wenn ein falsches Datenerhebungsinstrument ausgewählt wurde. Bei Stichproben sollte die Gesamtmenge so präzise wie möglich widerspiegeln, um die Datenakkuratesse nicht zu gefährden.[54] Weiterhin können technische und menschliche Fehler zu ungenauen Daten führen, indem Personen beispielsweise ihre persönlichen Informationen bei einer Befragung absichtlich verschleiern (z.B. Standardwert für Geburtsdatum 1. Januar), wobei man diese Problematik auch als *„disguised missing data“* bezeichnet.[55,56,57] Neben der falschen subjektiven Einschätzung des Menschen bei der Erhebung, können auch aus technischer Sicht ungenaue Daten ermittelt werden, wie z.B. durch (teils-)defekte Sensoren. Nicht zuletzt können Daten bei einem Transfer verfälscht werden bzw. sogar teilweise verloren gehen.[58]

Fehlenden Daten lassen sich einerseits durch technische Mängel begründen, andererseits auch durch die Tatsache, dass bestimmte Attribute schlichtweg von Beginn aus bei der Erhebung nicht betrachtet wurden oder durch bestimmte Restriktionen nicht verfügbar sind.[59]

Die aufzeigten Beispiele spiegeln nur einen kleinen Teil möglicher Ursachen wider und sollen die Bedeutsamkeit dieser Phase für den Data Mining Prozess aufzeigen. Die Datenvorbereitung stellt dabei einige mächtige Werkzeuge zur Verfügung, um die Datenqualität deutlich zu verbessern:[60,61,62]

- **Data Cleaning**: In diesem Schritt werden die Daten bereinigt, indem beispielsweise *fehlerhafte* oder *störende* Daten korrigiert werden. (siehe Kapitel 2.2.2.1 auf S. 15)

- **Data Integration**: Diese Phase beschäftigt sich mit der fehlerfreien Zusammenführung von Daten, da diese oftmals aus mehreren unterschiedlichen Quellen stammen. (siehe Kapitel 2.2.2.2 auf S. 16)

[52] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.84.
[53] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.196.
[54] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S.25.
[55] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.84.
[56] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S.24.
[57] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.196.
[58] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.84.
[59] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.84-85.
[60] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.11 ff..
[61] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.196 ff..
[62] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.84 ff..

- **Data Reduction**: Um die Algorithmen der Data Mining Methoden nutzen zu können, muss die immense Datenmenge reduziert bzw. komprimiert werden, um lange Laufzeiten zu vermeiden. (siehe Kapitel 2.2.2.3 auf S. 16)

Auf die in Abbildung 4 vereinfacht, dargestellten Werkzeuge und ihre Konzepte, wird im Folgenden näher eingegangen.

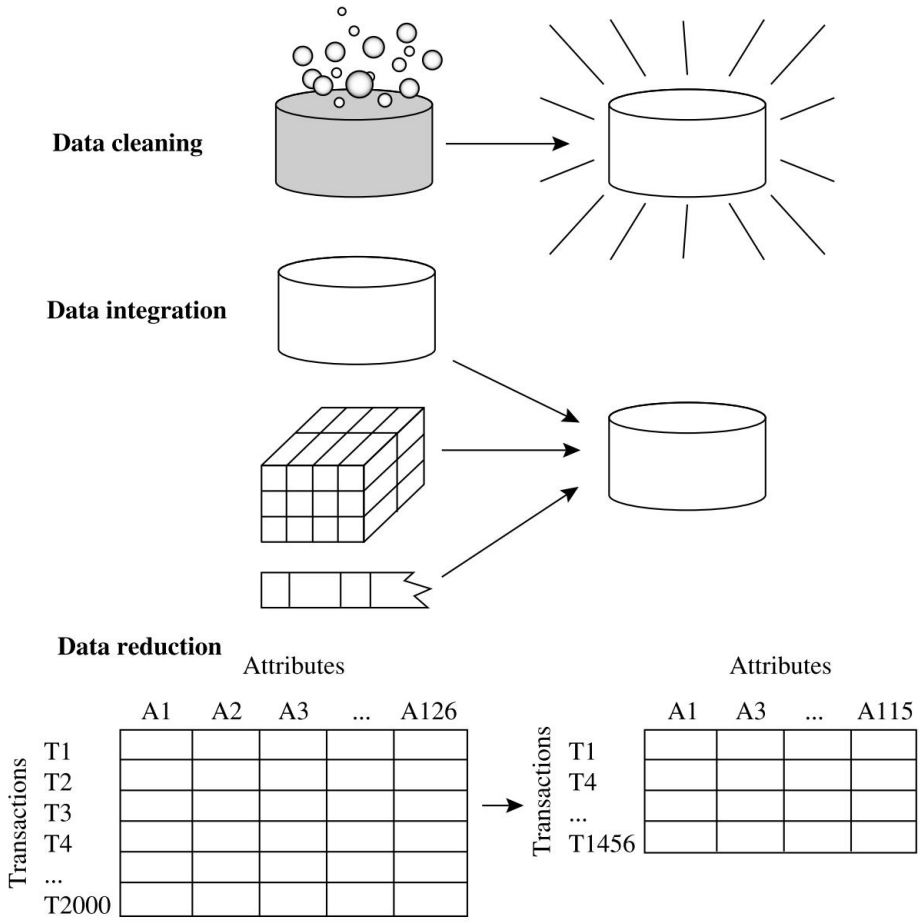

Abbildung 4: Werkzeuge der Datenvorverarbeitung[63]

[63] Vgl. Abbildung *Han*, Data Mining: Concepts and techniques, 2012, S.87

2.2.2.1 Data Cleaning

In der realen Welt sind Daten häufig „unvollständig, mit Fehlern oder Ausreißer behaftet oder sogar inkonsistent.“[64] Um Fehler oder gar falsche Resultate im Data-Mining-Prozess frühzeitig zu vermeiden, ist es von großer Bedeutung die Datenmenge zu bereinigen. Der Fokus sollte auf der Informationsneutralität liegen, das heißt, es sollen möglichst keine neuen Informationen hinzugefügt werden, die das reale Abbild verzerren oder verfälschen könnten.[65] Folgende Problemarten gilt es zu behandeln:[66,67]

Fehlende Daten Dem Datenanalyst stehen einige Möglichkeiten zur Verfügung, um auf fehlende Daten zu reagieren:

- *Attribut ignorieren*
  Der Datensatz mit dem fehlenden Attribut wird gänzlich ignoriert oder gelöscht. Jedoch können dadurch wichtige Informationen für die Datenanalyse verloren gehen, wodurch dieses Verfahren vor allem bei Datensätzen mit mehreren Lücken angewandt werden sollte.

- *Manuelles Einfügen*
  Besitzt der Datenanalyst das nötige Wissen, kann dieser einzelne Datensätze nachträglich manuell einfügen. Dieser Vorgang entwickelt sich schnell zu einem sehr zeitintensiven und unrealistischen Vorgang, der meistens undurchführbar ist, sobald die Datenmenge wächst. (z.B. 500 Kundendaten per Hand nachtragen)

- *Globale Konstante*
  Den fehlenden Wert durch eine globale Konstante ersetzen, ist sinnvoll, wenn auch ein leeres Feld als Information angesehen wird. Beispiele für Konstanten wären UNBEKANNT oder MINUS UNENDLICH.

- *Durchschnittswert*
  Handelt sich es bei dem fehlenden Attribut um einen metrischen Wert, so kann der Durchschnittswert aller Einträge als Ersatz verwendet werden. Der Durchschnittswert zeigt sich als äußerst einfache Möglichkeit, wenn die Daten klassifiziert werden können und dadurch Durchschnittswertberechnung nur auf

[64] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.199.
[65] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.199-200.
[66] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S.88-90..
[67] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S.200-201.

die Datensätze derselben Klasse angewandt wird. Die Methode der *k-Nearest Neighbours* steht zur Verfügung, wenn keine Klassen vorhanden sind, wobei der Durchschnitt, der dem aktuellen Datensatz ähnlichsten Werte benutzt wird.

- *Wahrscheinlichster oder häufigster Wert*
  Durch statistische Methoden kann der wahrscheinlichste Wert für das Attributs ermittelt werden, jedoch sollte dieser Ersatz begründet sein. Bei nicht numerischen Werten kann als weitere Möglichkeit auch der häufigste Wert, als Ersatz für das fehlende Attribut verwendet werden.

Verrauschte Daten und Ausreißer

- *Klasseneinteilung (bining)*
- *Regression*
- *Verbundbildung (clustering)*
- *Kombinierte Maschine/Mensch-Untersuchung*

Falsche und inkonsistente Daten

2.2.2.2 Data Integration

2.2.2.3 Data Reduction

2.2.3 Datentransformation

test

2.2.4 Data Mining Methoden

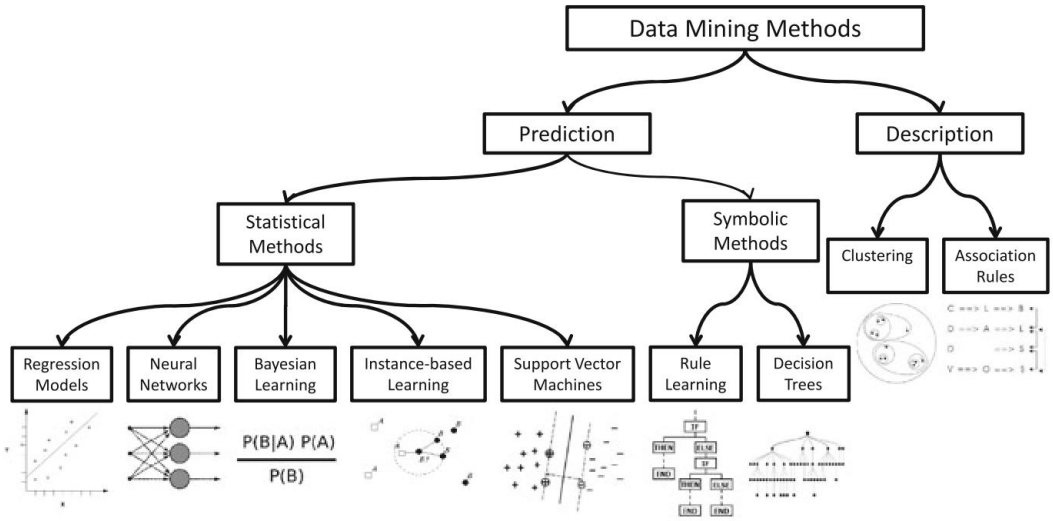

Abbildung 5: Übersichten Data Mining Methoden[68]

2.2.5 Interpretation

2.3 Funktionsmodellierung

2.3.1 Regressionsanalyse

2.3.2 MatLab

[68] Vgl. Abbildung *García* et al., Data preprocessing in data mining, 2015, S.4

3 Analysephase

3.1 Expected Goals

3.2 Opta-Spieldaten

4 Umsetzung

4.1 Datenselektion

4.2 Datenaufbereitung

4.3 Datentransformation

4.4 Modellierung der Funktion

4.4.1 Betrachtung des Winkels

4.4.2 Betrachtung der Distanz

4.4.3 Betrachtung der Koordinaten

4.5 Interpretation der Ergebnisse

5 Zusammenfassung

5.1 Fazit

5.2 Ausblick

A Annahmen

B MatLab Code

Literaturverzeichnis

Adhikari, Animesh/Adhikari, Jhimli [, 2015]: Advances in Knowledge Discovery in Databases. Band 79, Intelligent Systems Reference Library. Cham and s.l.: Springer International Publishing, 2015 ⟨URL: http://dx.doi.org/10.1007/978-3-319-13212-9⟩, ISBN 9783319132112

Aggarwal, Charu C. [, 2015]: Data mining: The textbook. Cham: Springer, 2015 ⟨URL: http://dx.doi.org/10.1007/978-3-319-14142-8⟩, ISBN 978–3–319–14142–8

Cleve, Jürgen/Lämmel, Uwe [, 2014]: Data Mining. [Elektronische Ressource] Auflage. München: De Gruyter Oldenbourg, 2014, ISBN 9783486713916

Fahrmeir, Ludwig et al. [, 2007]: Statistik: Der Weg zur Datenanalyse. 6. Auflage. Berlin: Springer, 2007, Springer-Lehrbuch ⟨URL: http://dx.doi.org/10.1007/978-3-540-69739-8⟩, ISBN 978–3–540–69739–8

Fasel, Daniel/Meier, Andreas (Hrsg.) [, 2016]: Big Data: Grundlagen, Systeme und Nutzungspotenziale. Wiesbaden: Springer Vieweg, 2016, Edition HMD ⟨URL: http://dx.doi.org/10.1007/978-3-658-11589-0⟩, ISBN 9783658115883

Fayyad, Usama/Piatetsky-Shapiro, Gregory/Smyth, Padhraic [AI Magazine 17 [1996], 1996]: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17 1996, Nr. 3, 37 ⟨URL: http://www.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131⟩, ISSN 0738–4602

García, Salvador/Luengo, Julián/Herrera, Francisco [, 2015]: Data preprocessing in data mining. Band 72, Intelligent Systems Reference Library. Cham: Springer, 2015 ⟨URL: http://dx.doi.org/10.1007/978-3-319-10247-4⟩, ISBN 978–3–319–10247–4

Han, Jiawei/Kamber, Micheline/Pei, Jian [, 2012]: Data mining: Concepts and techniques. 3. Auflage. Amsterdam: Elsevier/Morgan Kaufmann, 2012, The

Morgan Kaufmann series in data management systems ⟨URL: http://site.ebrary.com/lib/hamburg/Doc?id=10483440⟩, ISBN 978–0–12–381479–1

Mariscal, Gonzalo/Marbán, Óscar/Fernández, Covadonga [A survey of data mining processes, 2010]: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25 2010, Nr. 02, 137–166, ISSN 0269–8889

Runkler, Thomas A. [, 2015]: Data Mining: Modelle und Algorithmen intelligenter Datenanalyse. 2. Auflage. Wiesbaden: Springer Vieweg, 2015, Computational Intelligence ⟨URL: http://dx.doi.org/10.1007/978-3-8348-2171-3⟩, ISBN 978–3–8348–2171–3

Shi, Yong et al. [, 2015]: Intelligent knowledge: A study beyond data mining. s.l.: Springer-Verlag, 2015, SpringerBriefs in Business

Witten, Ian H./Frank, Eibe/Hall, Mark A. [Data mining: machine learning and techniques, 2011]: Data mining: Practical machine learning tools and techniques. 3. Auflage. San Francisco, Calif.: Kaufmann, 2011, The Morgan Kaufmann series in data management systems, ISBN 978–0–12–3748560

Ehrenwörtliche Erklärung

Ich versichere hiermit

- dass ich meine Bachelorarbeit mit dem Thema:

  Modellierung einer Funktion zur Berechnung der Wahrscheinlichkeit eines Torerfolges im Fußball

  selbstständig verfasst und
- keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.
- Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Ort, Datum Unterschrift