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
KNN	K-Nearest Neighbours
MATLAB	MATrix LABoratory
ML	Machine Learning
NN	Neuronale Netze
OLAP	Online Analytical Processing
SAP	eigenständiger Markenname - früher: *Systeme, Anwendungen und Produkte in der Datenverarbeitung*
SQL	Structured Query Language
SVM	Super Vector Machine

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 Markenname - 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 Unterstützung 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 folgende Kapitel soll dem Leser dazu eine Einführung in die Thematik geben, um ein Verständnis der grundlegenden Begrifflichkeiten und Ziele des Data Minings zu erlangen (vgl. Kapitel 2.1.1). Darüber hinaus werden die Prozesse des Data Minings (vgl. 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 beleuchtet wird.

2.1.1 Definition des Data Minings

Der Begriff des Data Minings reicht zurück bis in die 80er Jahre des letzten Jahrhunderts und verfolgt das Ziel, Wissen aus riesigen Datenmengen zu extrahieren.[3] Es ist ein Prozess des „*Sammelns, Säuberns, Verarbeitens und Analysierens von Daten, zur Gewinnung von nützlichen Informationen.*“[4] Denn der immense Datenanstieg in den letzten Jahrzehnten erlaubt uns nicht, einfach wertvolle Informationen oder organisiertes Wissen automatisch zu verstehen oder zu entnehmen. Erst das heutige „Informationszeitalter“ führte zum Beginn des renommierten Wissenschaftsbereich des Data Minings, welcher in der Literatur auch als natürliche Evolution der Informationstechnologie bezeichnet wird.[5,6] Grundlegende interdisziplinäre, wissenschaftliche Teilgebiete des Data Minings sind, z.B. die Statistik, das maschinelles Lernen (*Machine Learning*$^{\text{GL}}$ (*ML*)), die Mustererkennung, die Systemtheorie oder die *Künstliche Intelligenz*$^{\text{GL}}$.[7,8]

[3] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen, 2015, S. 2.
[4] *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, 2015, S. 2.
[8] Vgl. *Shi* et al., Intelligent knowledge, 2015, S. 1.

Cleve und Han vergleichen die Suche nach Mustern und Zusammenhängen in den Daten mit dem Abbau von Rohstoffen.[9] Sowie im Bergbau nach Schätzen wie Gold und Silber im Gestein gesucht wird, so strebt das *Data Mining*^{GL} (*DM*) nach dem Ableiten von Wissen aus den (Roh-)Daten.[10,11] Han geht sogar einen Schritt weiter und präferiert den Begriff des *Knowledge Mining from Data* – bezogen 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 bloße Betrachtung der Daten nicht aufgedeckt werden können. 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 wurde. 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.

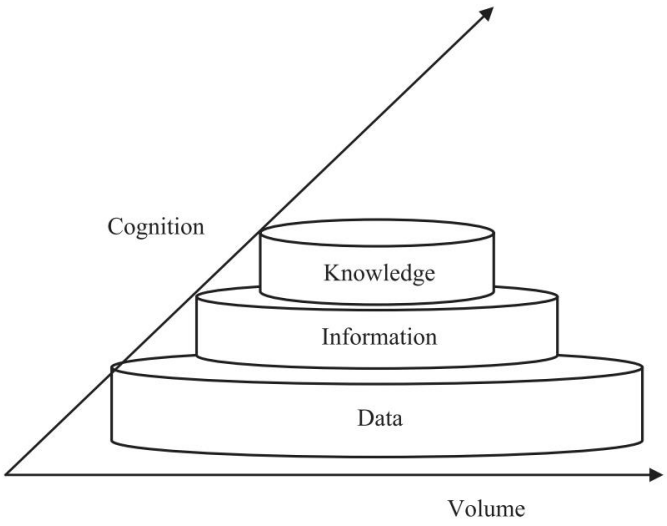

Abbildung 1: Wissensextraktion aus Daten[15]

[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. ebd.
[13] Weitere Termini nach Han: *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.*
[14] *Runkler*, Data Mining: Modelle und Algorithmen, 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 Datenmengen zu erheben, zu verarbeiten und zu analysieren, wodurch in diesem Kontext der Begriff *Big Data*[GL] entstanden ist.[16] Big Data bezeichnet Datenmengen, die mit herkömmlichen Analysemethoden nicht mehr zu verarbeiten wären und deshalb die Anwendung 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 entstandenen 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*[GL] (*IoT*). Einer der Trends der heutigen Informationstechnologie, welcher 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.[20,21,22]

Wir befinden uns in einer Welt, in der wir reich an Daten sind, jedoch arm an Informationen und Wissen. Der unglaublich rasante und gigantische Datenzuwachs hat bei weitem unsere menschliche Vorstellungskraft und Möglichkeiten übertroffen, sodass wir auf effiziente Werkzeuge angewiesen sind (siehe Kapitel 2.2.4 auf S. 23). Die sich immer weiter ausbreitende Spalte zwischen Daten und Informationen, führt nur noch durch die Nutzung von Methoden des Data Minings zu den begehrten

[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, 2015, S. 1.
[19] Vgl. *Aggarwal*, Data mining: The textbook, 2015, S. 2.
[20] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 39.
[21] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 1-2.
[22] Vgl. *Chu*, Data mining and knowledge discovery for big data, 2014, S. 85 ff.

„*Golden Nuggets of Knowledge*“.[23] Dazu müssen die (Roh-)Daten gezielt ausgewählt und umstrukturiert werden, um diese anschließend durch 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.“[24] Das Software-Tool *MatLab* wird innerhalb der Funktionsmodellierung in Kapitel 2.3 auf S. 27 vorgestellt und anschließend als Werkzeug zur Nutzung von Data Mining Methoden in der Umsetzungsphase genutzt (vgl. Kapitel 4 auf S. 30).

2.1.2 Data Mining Prozesse

In der Literatur grenzen viele Wissenschaftler den Begriff des eigentlichen Data Minings, vom Gesamtprozess der Extraktion von Wissen ab. Andere wiederum behandeln beide Termini synonym zu einander.[25,26,27] 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 adaptierter Form für die Methoden des Data Minings bereitstellen.[28] Hierzu werden im Folgenden kurz die zwei bekanntesten Prozessmodelle vorgestellt:

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

2.1.2.1 Knowledge Discovery in Data

Der Begriff des *Knowledge-Discovery-in-Data*-Prozesses wurde in den frühen 90-er Jahren geprägt und wird als „nicht trivialer Prozess zur Identifizierung von gülti-

[23] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 5.
[24] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen, 2015, S. 3.
[25] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 39.
[26] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 2.
[27] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 1.
[28] Vgl. ebd., 2015, S. 10.

gen, neuartigen, potentiell sinnvolle und letztlich verständlichen Muster in Daten“[29] definiert.[30] Erstmals wurde der Terminus von Gregory Piatetsky-Shapiro auf der *International Joint Conference on Artificial Intelligence*, 1989 in Detroit (USA), der Öffentlichkeit präsentiert.[31] Der in Abbildung 2 auf S. 9 dargestellte iterative KDD-Prozess nach Fayyad, beinhaltet folgende Schritte, wobei das DM als ein eigener Prozessschritt ausgewiesen wird:[32]

1. **Datenselektion**: Auswahl der geeigneten Datenmengen.

2. **Datenvorverarbeitung**: Behandlung fehlender oder problembehafteter Daten.

3. **Datentransformation**: Umwandlung in adäquate Datenformate.

4. **Data Mining**: Suche nach Muster.

5. **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. Die Abkürzung *KDD* steht in der Literatur für unterschiedliche Bezeichnungen, wie zum Beispiel *Knowledge Discovery in Databases*, *Knowledge Discovery in Data Mining* oder *Knowledge Discovery in Data Warehouses*.[33] Alle zielen dabei auf die Erforschung von Wissen aus Datenmengen ab, wodurch in dieser Arbeit die allgemeingültige Bezeichnung *Knowledge Discovery in Data* verwendet wird.

[29] *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 41.
[30] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 2.
[31] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S. 1.
[32] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 5.
[33] Vgl. *Osei-Bryson/Barclay*, Knowledge discovery process and methods, 2015, S. 26 ff.

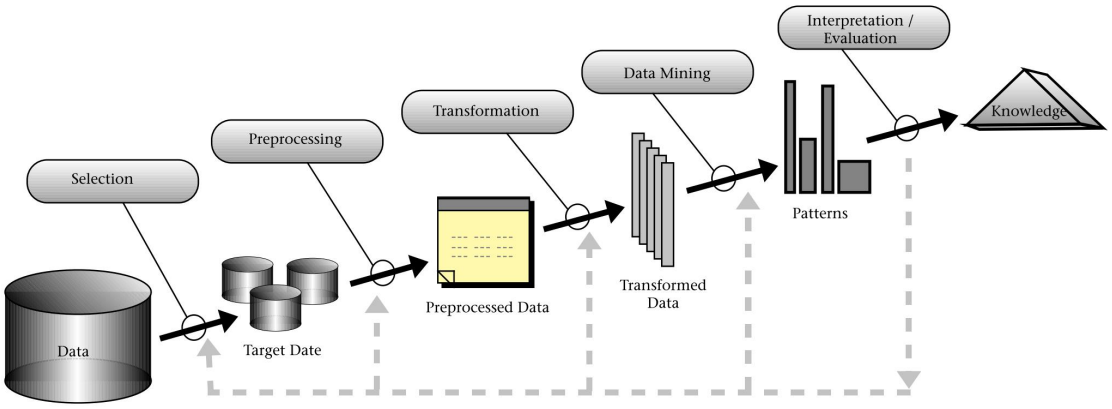

Abbildung 2: Der Knowledge Discovery in Data Prozess[34]

2.1.2.2 CRISP-DM

Das CRISP-DM-Modell wurde im Jahr 2000 durch ein Konsortium, bestehend aus mehreren Firmen, entwickelt. Beteiligt daran waren:[35,36]

- 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 sechs beinhaltenden Etappen ausgegangen, der in Abbildung 3 auf S. 10 dargestellt werden.[37]

[34] Vgl. Abbildung *Fayyad* et al., From Data Mining to Knowledge, 1996, S. 41.
[35] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 6.
[36] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 3.
[37] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 6-8.

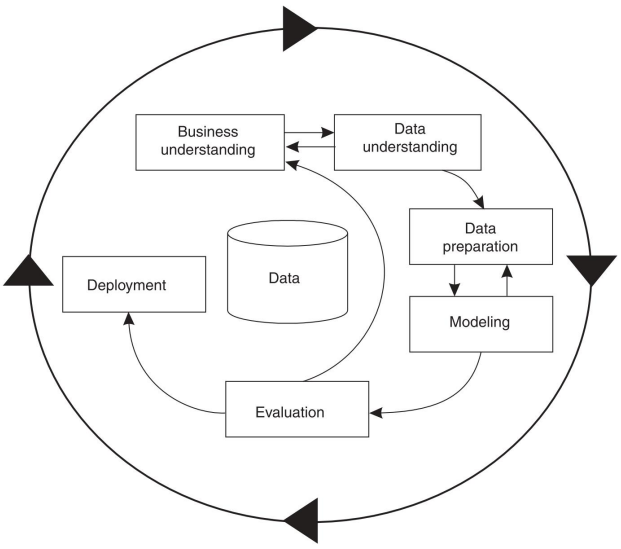

Abbildung 3: CRISP-DM Prozess[38]

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

2. Verständnis der Daten: Diese Phase beschäftigt sich mit den benötigten Daten zur Durchführung der Analyse. Daten werden gesammelt und beschrieben, um deren betriebliche Bedeutung zu verstehen.

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

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

[38] Vgl. Abbildung *Mariscal* et al., A survey of data mining, 2010, S. 13

5. Evalution: Die erzielten Ergebnisse werden an den aus Phase 1 definierten Erfolgskriterien gemessen, um beispielsweise festzustellen, ob der wirtschaftliche Nutzen erzielt wurde.

6. Einsatz im Unternehmen: Zuletzt gilt es den Einsatz der Resultate in das Unternehmen vorzubereiten und in das operative Geschäft zu integrieren.

Das Modell bezieht und orientiert sich, wie schon am Namen zu erkennen ist, stark an wirtschaftlichen Projekten und beschreibt *Was* zu tun ist, jedoch nicht genau *Wie*, sodass Projektteams innerhalb dieses Rahmens beginnen ihre eigenen Methoden zu verwenden.[39]

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.[40] Im Gegensatz dazu konzentriert sich der KDD-Prozess auf die Datenbereitstellung und Analyse, sodass dieser als grundlegende Methodik für die spätere Umsetzung der wissenschaftlichen Aufgabenstellung herangezogen wird und genauer in Kapitel 2.2 beleuchtet 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.[41] Nicht zuletzt ist zu erwähnen, dass in der Literatur unterschiedliche Auffassungen zu dem Begriff des Data Minings 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

[39] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 4.
[40] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 8.
[41] Vgl. vorgestellte Modelle aus *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010.

Arbeit ausgewählt wurde. Hierzu werden die einzelnen Prozessschritte der Datenselektion, der Datenvorverarbeitung, der Datentransformation, der Data-Mining-Methoden, sowie der Interpretation der Ergebnisse konkretisiert, um diese in der späteren Umsetzung der wissenschaftlichen Aufgabe anwenden zu können.

„Experten [...] haben realisiert, dass eine 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*[GL] (*SQL*)-Abfrage oder *Online Analytical Processing*[GL] (*OLAP*) reichen für eine komplexe Datenanalyse oft nicht aus.“[42] Hier greift der in Abbildung 2 auf S. 9 dargestellte KDD-Prozess, ein multiples iteratives Modell, indem die einzelnen Schritte solange wiederholt und aufeinander abgestimmt werden müssen, bis aus den zugrundeliegenden Daten, Wissen abgeleitet werden kann.[43] 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. 27) – an riesige Datenmengen genutzt werden.[44,45]

In der Literatur existieren unterschiedliche Vorstellungen der einzelnen Prozessschritte, wodurch es oftmals zu Überschneidungen zwischen den einzelnen Gebieten kommt. So findet sich die Methode der *Data Integration* einerseits in der Datenselektion wieder, andererseits auch in der Datenvorverarbeitung.[46,47] Im Folgenden wird versucht, diese Schritte klar von einander abzutrennen. Hierbei wird sich größtenteils an den Ausarbeitungen von Han et al. und Cleve et al. orientiert.

2.2.1 Datenselektion

Die Datenselektion befasst sich hauptsächlich mit der Auswahl der geeigneten Datenmengen – der *Zieldaten* – auf Basis derer die spätere Erforschung ausgeübt wird.[48] Der Datenanalyst befasst sich in dieser Phase mit der Bestimmung der für die Analyse geeigneten Daten und des Exports dieser Datenauswahl beispielsweise in eine Datenbank. Die selektierten Daten können zum Beispiel technischen oder rechtlichen

[42] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S. 1.
[43] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 7.
[44] Vgl. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S. 1.
[45] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 7.
[46] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 1.
[47] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 198.
[48] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 42.

Restriktionen unterliegen, wie zum Beispiel Zugriffs- oder Kapazitätsbeschränkungen. Hierbei sollte auf eine repräsentative Teilmenge des Datenbestandes zurückgegriffen werden.[49]

2.2.2 Datenvorverarbeitung

„Da die Zieldaten aus den Datenquellen lediglich extrahiert werden, ist im Rahmen der Datenvorverarbeitung die Qualität des Zieldatenbestandes zu untersuchen und – sofern nötig – dieser durch den Einsatz geeigneter Verfahren zu verbessern.“[50]

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 bereitstellen zu können. Die Struktur und das Format müssen perfekt auf die vorliegende Aufgabe passen, ansonsten führt die geringe Qualität der Daten zu schlechten bzw. falschen Resultaten, bis hin zu Laufzeitfehlern.[51] Es gilt auch hier das alte Prinzip: GIGO – garbage in, garbage out.[52] Die oftmals schlechte Qualität der (Roh-)Daten ist durch *fehlende, ungenaue, inkonsistente bzw. widersprüchliche* Daten zu begründen.[53,54] Im Folgenden werden dazu einige Ursachen beispielhaft 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 widergespiegelt werden, um die Datenakkuratesse nicht zu gefährden.[55] 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.[56,57,58] Neben der falschen subjektiven Einschätzung des Menschen bei der Erhebung, können auch aus technischem Blickpunkt ungenaue Daten ermittelt werden, wie z.B. durch (teils-)defekte Sensoren. Nicht zuletzt können Daten bei

[49] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 9.
[50] Ebd.
[51] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 10-11.
[52] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 197.
[53] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 84.
[54] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 196.
[55] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 25.
[56] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 84.
[57] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 24.
[58] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 196.

einem Transfer verfälscht werden bzw. sogar teilweise verloren gehen.[59]

Fehlende Daten lassen sich einerseits durch technische Mängel begründen, andererseits auch durch die Tatsache, dass bestimmte Attribute schlichtweg von Beginn aus bei der Erhebung nicht beachtet wurden oder durch bestimmte Restriktionen nicht verfügbar sind.[60]

Die aufgezeigten 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 nachhaltig zu verbessern:[61,62,63]

- **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. 18).

- **Data Reduction**: Um die Algorithmen der Data Mining Methoden nutzen zu können, muss die immense Datenmenge reduziert bzw. komprimiert werden, um lange Laufzeiten zu verhindern (siehe Kapitel 2.2.2.3 auf S. 20).

Auf die in Abbildung 4 auf S. 15 vereinfacht, dargestellten Werkzeuge und ihre Konzepte, wird in den folgenden Unterkapitel näher eingegangen.

[59] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 84.
[60] Vgl. ebd., 2012, S. 84-85.
[61] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 11 ff.
[62] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 196 ff.
[63] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 84 ff.

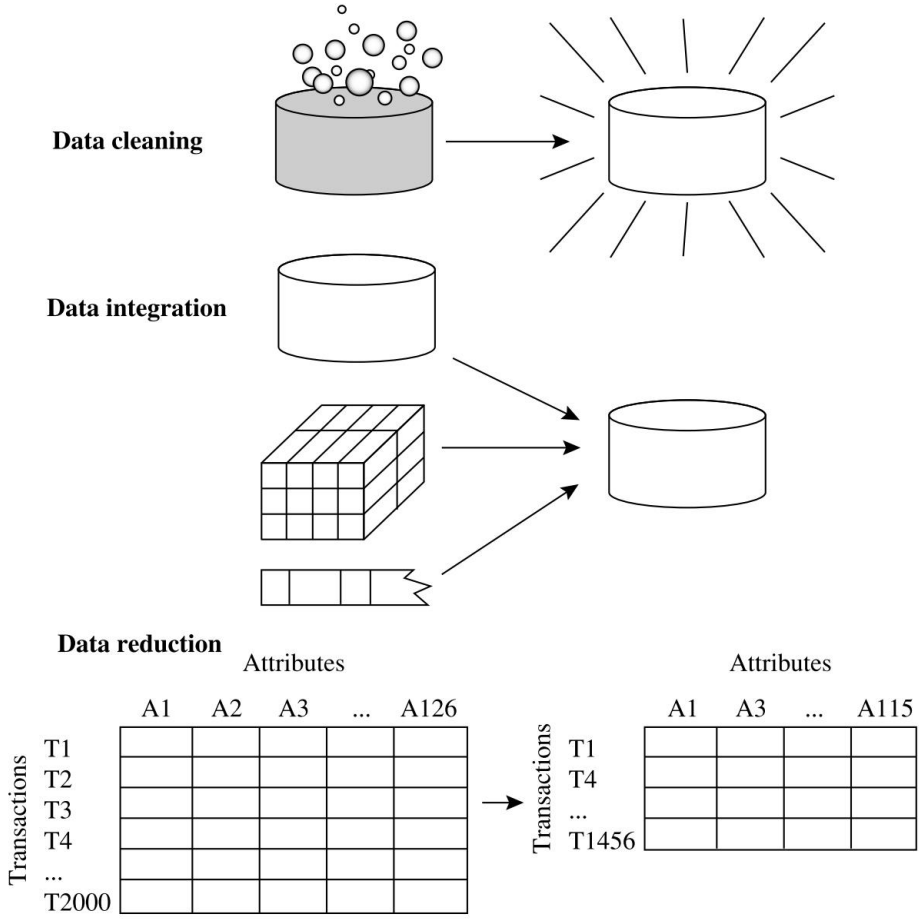

Abbildung 4: Werkzeuge der Datenvorverarbeitung[64]

2.2.2.1 Data Cleaning

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

[64] Vgl. Abbildung *Han*, Data Mining: Concepts and techniques, 2012, S. 87
[65] *Cleve/Lämmel*, Data Mining, 2014, S. 199.
[66] Vgl. ebd., 2014, S. 199-200.

Fehlende Daten Dem Datenanalyst stehen einige Möglichkeiten zur Verfügung, um auf fehlende Daten reagieren zu können:[67,68]

- *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 nur 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 zeitaufwändigen und unrealistischen Vorgang, der aufgrund des Mangels an Ressourcen (personeller wie auch zeitlicher) undurchführbar ist, sobald die Datenmenge wächst (z.B. 500 Kundendaten per Hand nachtragen).

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

- *Durchschnittswert*
  Handelt es sich 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 die Durchschnittswertberechnung nur auf Datensätzen der selben Klasse angewandt wird. Die Methode der *K-Nearest Neighbours*$^{\text{GL}}$ (*KNN*)[69] steht zur Verfügung, wenn keine Klassen vorhanden sind. Hierbei wird der Durchschnitt, der dem aktuellen Datensatz ähnlichsten Werte benutzt wird.

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

[67] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 88-90.
[68] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 200-201.
[69] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 76.

Verrauschte Daten und Ausreißer Durch ungenaue Messwerte oder falschen Schätzungen entstehen die sogenannten *verrauschten Daten*.[70] Um diese bereinigen zu können, stehen dem Datenanalyst einige Verfahren zur Verfügung, wodurch diese fehlerbehafteten Daten angeglichen werden können.[71] Als *Ausreißer* bezeichnet man dabei Daten, die erheblich von den anderen Daten abweichen oder außerhalb eines Wertebereiches liegen.[72] Beispielsweise liegen Daten von 30- bis 50- Jährigen vor, darunter auch einer von einem 90-Jährigen. Hierbei könnte es sich um einen Ausreißer handeln, aber auch um einen fehlerhaften Datensatz.[73] „Ob solche Ausreißer für das Data Mining ausgeblendet oder adaptiert werden sollten oder besser doch im Originalzustand zu verwenden sind, hängt vom konkreten Kontext ab.“[74]

- *Klasseneinteilung (bining)*
  Durch die Gruppierung verrauschter Daten in Klassen, können diese beispielsweise durch den Mittelwert oder die naheliegenden Grenzwerte ersetzt werden.

- *Regression*$^{\text{GL}}$
  Die Darstellung der Daten in Form einer mathematischen Funktion, bietet die Möglichkeit, fehlerbehaftete Daten durch die berechneten Funktionswerte zu ersetzen. Für zwei Abhängigkeiten zwischen zwei Attributen steht hierbei neben der *linearen Regression*, auch die *multiple lineare Regression* für mehrere Attribute als Werkzeuge zur Verfügung (weiterführende Ausarbeitung zur Regressionsanalyse siehe Kapitel 2.3.1 auf S. 28).

- *Verbundbildung (clustering)*
  Eine der einfachsten Möglichkeiten um Ausreißer zu erkennen, bietet die Verbundbildung, auch *Clustering*$^{\text{GL}}$ genannt. Hierbei werden ähnliche Daten, wie in Abbildung 5 auf S. 18 dargestellt, zu *Clustern* zusammengeführt, wodurch sich die Ausreißer direkt identifizieren lassen.

[70] Im englischen Sprachgebrauch als *noisy data* bekannt.
[71] Auch als *smoothing* bekannt.
[72] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 89-90.
[73] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 196.
[74] Ebd.

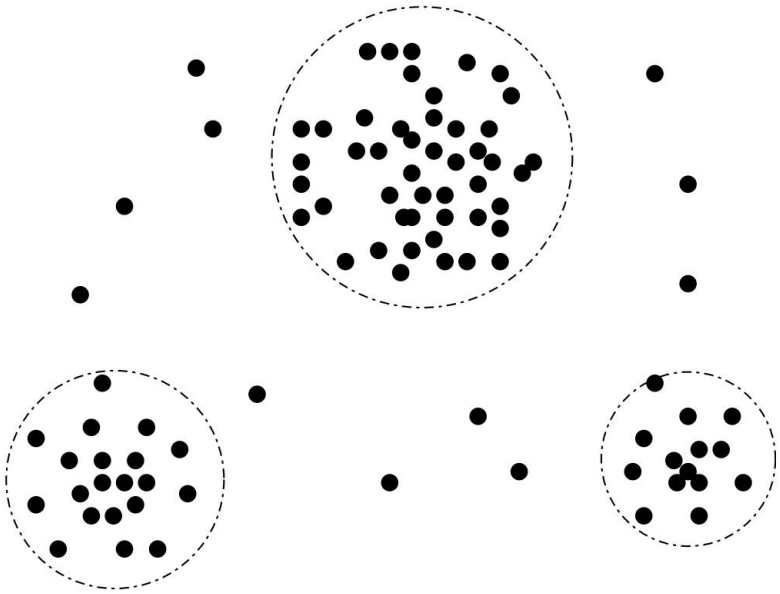

Abbildung 5: Outlierdetection mittels Clustering[75]

Falsche und inkonsistente Daten Bei falschen bzw. inkonsistenten Daten ergeben sich prinzipiell zwei Möglichkeiten zur Korrekturbehandlung. Einerseits können der Datensatz oder bestimmte Attribute durch *Löschen* entfernt werden, wobei jedoch die Gefahr einer zu großen Reduktion des Datenbestandes entsteht und relevante Informationen für das Data Mining verloren gehen könnten. Die zweite Korrekturvariante versucht den inkonsistenten Datensatz, durch die *Zuhilfenahme anderer Datensätze*, sinnvoll zu ersetzen. Sollte eine Unterscheidung zwischen *falsch* und *richtig* nicht möglich sein, wären beim Löschen immer mindestens zwei Datensätze betroffen.[76]

2.2.2.2 Data Integration

Bei Data-Mining-Projekten ist oftmals die Integration mehrerer Datenbestände aus unterschiedlichen Quellen erforderlich. Diese Phase sollte mit äußerster Sorgfalt durchgeführt werden, um frühzeitig redundante und inkonsistente Datensätze zu vermeiden, wodurch die Genauigkeit und Geschwindigkeit der nachfolgenden Data Mining Algorithmen nicht gefährdet wird.[77] Folgende Punkte gilt es bei der Daten-

[75] Vgl. Abbildung *Han*, Data Mining: Concepts and techniques, 2012, S. 91
[76] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 203-204.
[77] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 93-94.

integration zu beachten:

- **Identifikationsproblem von Entitäten**:
  Bei der Datenintegration aus multiplen Datenquellen, wie beispielsweise Datenbanken oder Dokumenten, stellt die Schema-Integration wie auch die Objektanpassungen eine schwierige Herausforderung dar. Der Datenanalyst muss sicherstellen, das zum Beispiel das Attribut *kunden_nummer* aus der einen Datenquelle, die selbe Referenz besitzt, wie das Attribut *kunden_id* aus einer anderen und es sich folglich um das selbe Attribut handelt. Dies wird allgemein als *Problem Identification Problem* bezeichnet.[78,79] Die Metadaten der Attribute beinhalten Informationen, wie *Name, Bedeutung, Datentyp, Wertebereich, uvm.* und können durch Abgleich derer zu einer hilfreichen Vermeidung von Fehlern bei der Integration beitragen. Weiterhin muss gesondert auf die *Datenstruktur* geachtet werden, um keine referentiellen Abhängigkeiten bzw. Beziehungen zwischen den Daten zu zerstören.[80]

- **Redundanzen bei Attributen**:
  Ein Attribut, welches durch ein anderes Attribut ableitbar ist – wie zum Beispiel das Alter vom Geburtsjahr berechnet werden kann – wird als redundant bezeichnet. Die Vielzahl von Redundanzen führt zu unnötig aufgeblähten Datenmengen, die wiederum die Performanz sowie die Resultate eines Data Mining Algorithmus negativ beeinträchtigen können.[81] Folglich sollte diese Problematik durch die Anwendung von statistischen Verfahren, in Form der Korrelationsanalyse, dezidiert behandelt werden. Für numerische Werte ist dabei der Einsatz von Korrelationskoeffizienten und Kovarianzen hilfreich. Um die Implikation zweier Attribute einer nominalen Datenmenge[82] bestimmen zu können, verwendet man in der Regel den χ^2(*Chi*^{2})-Test.[83,84,85]

- **Duplikatserkennung**:
  Duplikate verkörpern Redundanzen auf Datensatzebene und führen einerseits zu unnötig großen Datenmengen, die sich wiederum auf die Performanz der Algorithmen auswirken. Andererseits führt jedoch auch die verfälschte Gewichtung der mehrfach vorkommenden Datensätze, zu schlichtweg falschen

[78] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 199.
[79] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 94.
[80] Vgl. ebd.
[81] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S.41.
[82] Rein qualitative Merkmalsausprägungen ohne natürliche Rangordnung (wie z.B. das Geschlecht).
[83] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. ..
[84] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 41.
[85] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 64.

Analyseergebnissen. Ein häufiger Grund stellt dabei die Verwendung von denormalisierten Datenbanktabellen dar.[86,87]

- **Konflikte bei Attributswerten**:
  Hierbei handelt es sich um die unterschiedliche Darstellung, Skalierung und Kodierung von Attributswerten. Beispielsweise kann das Attribut *Gewicht* durch das metrische System oder das britische Maßsystem repräsentiert werden, woraus bei der Integration von Daten zu einer einheitlichen Quelle immer wieder Konflikten resultieren.[88,89]

2.2.2.3 Data Reduction

Die bereits mehrfach angesprochene Problematik der riesigen Datenmengen bei Data-Mining-Projekten, steigert die Komplexität und vermindert die Effizienz der Algorithmen. Daher strebt die Datenreduktion – wie die Bezeichnung erkennen lässt – nach einer reduzierten repräsentativen Teilmenge, welche die Integrität des Originals nicht verliert. Dazu können folgende drei Techniken angewandt werden:[90,91,92]

1. Dimensionsreduktion
2. Datenkompression
3. Numerische Datenreduktion

Dimensionsreduktion Hierbei bleiben irrelevante Attribute des Datensatzes unberücksichtigt und nur für die Analyse relevante Daten werden mit einbezogen. Allgemein empfehlen sich dafür zwei Verfahren: Bei der schrittweisen *Vorwärtsauswahl* werden wesentliche Attribute einer sukzessiv wachsenden Zielmenge zugeordnet. Im Gegensatz dazu werden bei der *Rückwärtseliminierung* die uninteressanten Daten schrittweise aus der Zielmenge eliminiert.[93]

[86] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 98.
[87] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 43.
[88] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 99.
[89] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 199.
[90] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 147 ff.
[91] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 99-100.
[92] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 206-208.
[93] Vgl. ebd., 2014, S. 206.

Datenkompression Bei dieser Technik wird durch Transformation oder Codierung versucht, eine Reduktion der Datenmenge zu erreichen. Fasst man beispielsweise die einzelnen Attribute *Tag, Monat und Jahr* zu einem neuen Attribut *Datum* zusammen, können Datensätze komprimiert werden.[94]

Numerische Datenreduktion Statt die gesamte Datenmenge für die Analyse heranzuziehen, wird innerhalb der numerischen Datenreduktion eine repräsentative Teilmenge – in Form einer Stichprobe – für das DM genutzt. Im Vordergrund steht hierbei die passende Auswahl unterschiedlicher Stichprobenverfahren, wie der *zufälligen Stichprobe* oder der *repräsentativen Stichprobe*, wobei kein verzerrtes Abbild der Daten resultieren darf.[95,96]

2.2.3 Datentransformation

Nachdem die (Roh-)Daten selektiert, bereinigt und auf eine relevante Zielmenge reduziert wurden, müssen diese nur noch in eine adaptierte Form für die Algorithmen des Data Minings transformiert werden.[97] Oftmals müssen sogar neue Attribute aus einem Datensatz kreiert werden, da diese nicht in geeigneter Struktur für das Data-Mining-Verfahren vorliegen.[98] Dazu es gibt eine Reihe an unterschiedlichen Transformationsmöglichkeiten, wobei in dieser Arbeit ein Auszug der relevanten Methoden vorgestellt werden soll:

Codierung Liegen beispielsweise Attribute mit einer ordinalen Ausprägung vor (wie *sehr groß, groß, mittel und klein*), müssen diese bei einer Verwendung des KNN-Algorithmus in numerische Werte umgewandelt werden (Werte zwischen 0 und 1). Hierbei würde sich folgende Codierung für das Attribut *Körpergröße* anbieten:[99]

- *sehr groß* $\rightarrow$ 1
- *groß* $\rightarrow$ 0,66
- *mittel* $\rightarrow$ 0,33

[94] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 207.
[95] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 25-27.
[96] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 207.
[97] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 112.
[98] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 48.
[99] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 210.

- *klein* $\rightarrow 0$

Die Ordnungsrelation, hier *sehr groß* > *groß* > ..., darf dabei jedoch nicht verloren gehen. In Abhängigkeit zu dem jeweiligen Verfahren, müssen Daten, sowie dies bei Maßeinheiten immer wieder der Fall ist, oftmals kodiert werden.[100]

Normalisierung und Skalierung Unterschiedliche Maßeinheiten – wie *Körpergröße* und *Körpergewicht* – können die Datenanalyse negativ beeinflussen und müssen daher in eine einheitliche Skalierung transformiert werden, um eine gleiche Gewichtung aller Attribute zu erreichen. Man bedient sich hierbei in der Regel an der *Min-Max-Normalisierung* (siehe Abbildung 6) oder der *Z-Transformation*, um numerische Werte auf ein [0,1] Intervall zu normieren.[101,102]

$$x_{neu} = \frac{x - min(x_i)}{max(x_i) - min(x_i)} \tag{1}$$

Abbildung 6: Min-Max-Normalisierung

Datenaggregation Nicht nur aus Sicht der Datenkompression (vgl. Kapitel 2.2.2.3 auf S. 21) ist die Datenaggregation erforderlich. Vielmehr „kann die Aggregation aus inhaltlichen Gründen sinnvoll sein.“[103] Wenn Daten auf einer zu detaillierten Ebene vorliegen – wie beispielsweise Einwohnerzahlen von Stadtteilen – müssen diese für einen Städtevergleich erst summiert werden, um bundesweite Aussagen treffen zu können. Je nach Kontext können verschiedene Aggregationsmethoden (wie z.B. Summenbildung, Durchschnitt, usw.) für die Transformation zu einem einzigen Wert angewendet werden.[104]

Datenglättung Die bereits in Kapitel 2.2.2.1 auf S. 16 vorgestellten Techniken zur Bereinigung von verrauschten Daten und Ausreißer, finden auch bei der Transformation ihre Verwendung. Die Datenglättung strebt nach einer reduzierten Datenmenge,

[100] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 211.
[101] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 114.
[102] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 212.
[103] Ebd., 2014, S. 214.
[104] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 112.

worin jeder numerische Wert durch idealisierte Werte, wie beispielsweise der *Regression*, ersetzt wird.[105]

2.2.4 Data-Mining-Methoden

Nachdem die Daten in geeigneter Form vorliegen, kommt das eigentliche Herzstück des KDD-Prozesses – das *Data Mining* – zum tragen. In diesem Schritt wird zu nächst festgestellt, welche grundlegende Data-Mining-Aufgabe es zu lösen gilt, um anschließend ein passendes Analyseverfahren zur Identifizierung von Mustern und Zusammenhänge auswählen zu können.[106] Die interdisziplinäre Wissenschaft des Data Minings umfasst bewährte Techniken aus vielen Forschungsgebieten, welche auf verschiedenste Problemfälle der Realität, wie Zeitreihenanalysen, Funktionsmodellierungen, Klassifikation uvm., angewendet werden können. Grundsätzlich basieren fast alle Analyseverfahren auf der Mathematik, insbesondere der Statistik.[107] Im allgemeinen unterscheidet man die Data-Mining-Methoden in zweit Kategorien: *Prognose* und *Beschreibung*. Hierzu gibt Abbildung 7 auf S. 26 einen guten Überblick über die Einteilung der etablierten Methoden, welche im Folgenden kurz aufgeführt werden.

Prognose: In dem Bereich der Prognose unterscheidet man zwischen zwei Gruppen: *statistische Methoden* und *symbolische Methoden*. Letztere versuchen das Wissen durch Symbolik und Verknüpfung, auf einer leichter interpretierbaren Ebene für Menschen, zu vermitteln. Im Gegensatz dazu, repräsentieren statistische Methoden das Wissen mit Hilfe der Berechnung von mathematischen Modellen.[108] Die am häufigst angewendeten statistischen Methoden sind:[109,110]

- *Regressionsanalyse*
  Die älteste DM-Methode dient zur Funktionsmodellierung von einer abhängigen oder mehreren unabhängigen Variablen. Die Form der Funktion wird dabei durch das ausgewählte Verfahren, beispielsweise *lineare oder quadratische Regression*, bestimmt und kann anhand bestimmter Parameter validiert werden, wie „gut“ diese zu den eingebrachten Daten passt.[111]

[105] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 214-215.
[106] Vgl. ebd., 2014, S. 10.
[107] Vgl. ebd., 2014, S. 12.
[108] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 3.
[109] Vgl. ebd., 2015, S. 3-5.
[110] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 23-24.
[111] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 3.

- *(Künstliche) Neuronale Netze*[GL] *(NN)*
  In diesem Teilbereich der Künstlichen Intelligenz wird versucht einen Wissensspeicher zu kreieren, der ähnlich unserem leistungsfähigen Gehirn funktionieren soll. Hierbei werden die biologischen Elemente und Vorgehensweise des Gehirns, in Form von *Neuronen*, in die Welt des Computers übertragen. Durch gerichtete und gewichtete Verbindungen sind diese Neuronen untereinander verknüpft und bilden so ein gemeinsames Netz für die Informationsverarbeitung.[112]

- *Super Vector Machine*[GL] *(SVM)*
  Die auf ML basierende Methode versucht Objekte zu klassifizieren. Dabei werden alle Objekte als Vektoren in einem Raum repräsentiert und durch sogenannte *Hyperebenen* (fungieren als Trennflächen) geteilt, um eine möglichst zuverlässige Zuordnung der Daten in vordefinierte Klassen zu erreichen.[113]

Im Bereich der symbolischen Methoden hat sich die Technik des *Entscheidungsbaumes* etabliert. Sie dient ebenfalls der Klassifizierung von Objekten, indem pro Iterationsschritt das am *besten* zu klassifizierende Attribut gefunden wird, um die Daten daran aufzusplitten. Durch dieses Verfahren entsteht ein Entscheidungsbaum, anhand dem Regeln, wie *If-Else-Zweige*, abgeleitet werden können.[114]

Beschreibung:

- *Clustering*
  Im Gegensatz zur Klassifizierung sind bei der Methode des Clustering zuvor keine Klassen bzw. Gruppen definiert. Dieses weitverbreitete Werkzeug im Bereich des Data Minings versucht Daten in sogenannte *Cluster* zu unterteilen, wobei die Elemente dieser Gruppe sich möglichst ähnlich (*homogen*), jedoch auch gleichzeitig von den anderen Clustern deutlich zu unterscheiden sein sollten (*heterogen*).[115]

- *Assoziationsanalyse*
  Diese Methode versucht Wissen durch assoziative Beziehungen zwischen den Daten herzuleiten. Das einfachste Beispiel hierfür wäre im Einzelhandelsbereich: „Wenn ein Kunde Produkt A kauft, würde dieser auch Produkt B kaufen." Durch diese extrahierten Muster, können wiederum Regeln abgeleitet

[112] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 47.
[113] Vgl. *Aggarwal*, Data mining: The textbook, 2015, S. 313.
[114] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 5.
[115] Vgl. *Anderberg*, Cluster Analysis for Applications, 2014, S. 3.

werden.

Nicht zuletzt ist die Visualisierung unerlässlich für den Erfolg eines Data-Mining-Projektes. Die Resultate werden oftmals zur Entscheidungsfindung herangezogen, wobei die Entscheidungsträger nicht immer direkt am Prozess beteiligt waren. Die Ergebnisse müssen folglich in einer anschaulichen und nachvollziehbaren Form dargestellt werden, um Vertrauen und Akzeptanz in die Resultate zu gewinnen.[117] Weiterhin kann die Visualisierung auch schon in der Datenvorverarbeitung genutzt werden oder als eigenständige Methode innerhalb des Data Minings, da sich häufig erst Zusammenhänge zwischen den Attributen durch die Darstellung der Daten erkennen lassen.[118]

Für die Modellierung einer Funktion zur Berechnung der Wahrscheinlichkeit eines Torerfolges im Fußball (auch bekannt unter dem Begriff *Expected Goals*), kann die passende Data-Mining-Methode aus der Abbildung 7 auf S. 26 ausgewählt werden. Der zu erwartende Torerfolg soll folglich prognostiziert und durch ein mathematisches Modell repräsentiert werden. Unter den statistischen Methoden eignet sich für die Modellierung einer Funktion am besten die Regressionsanalyse, da ein Torerfolg von mehreren Faktoren abhängig ist. Dementsprechend wird dieses Verfahren als Data-Mining-Methode für die Beantwortung der vorliegenden Problemstellung ausgewählt und dessen Bestandteile zunächst in Kapitel 2.3 auf S. 27 betrachtet, um diese Technik in der späteren Umsetzung anwenden zu können.

2.2.5 Interpretation

Am Ende jedes KDD-Prozesses steht die Interpretation sowie die Evolution der entdeckten Muster und Beziehungen aus dem Data Mining. Oftmals können Unternehmen keinen Nutzen aus den Analyseverfahren erzielen, da diese häufig irrelevante, triviale, bedeutungslose oder sogar bereits bekannte Daten generieren. Die gewonnenen Muster sollten den folgenden vier Kriterien genügen, um neues Wissen zu repräsentieren:[119]

1. **Validität:** Hierbei wird die Gültigkeit des Muster für das gefundene Modell, als auch in Bezug auf neue Daten, in einem objektiven Maßstab bewerten.

2. **Neuartigkeit:** Das Kriterium beantwortet die Frage, inwiefern das neu er-

[117] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 14.
[118] Vgl. ebd.
[119] Vgl. ebd., 2014, S. 11-12.

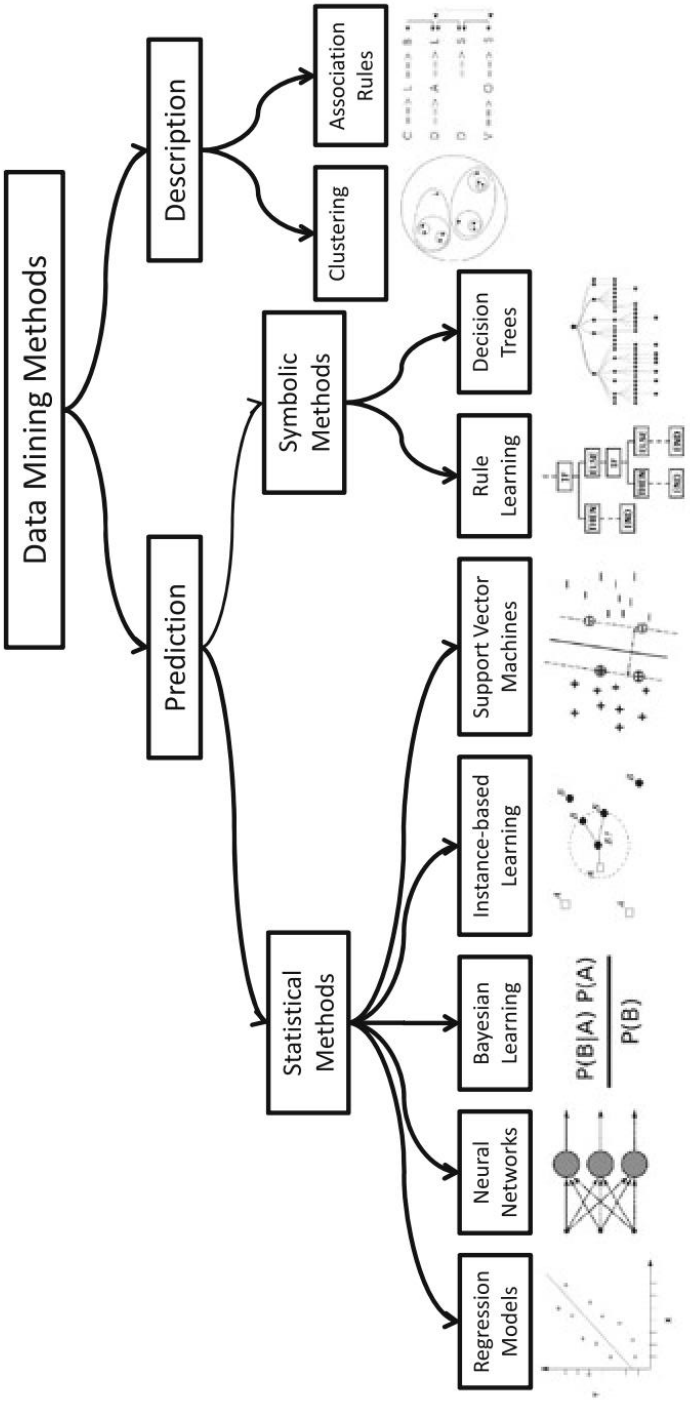

Abbildung 7: Übersicht: Data-Mining-Methoden[116]

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

worbene Wissen zu den bisherigen Forschungen steht. Einerseits es kann den Wissensstand ergänzen oder im Widerspruch dazu stehen.

3. **Nützlichkeit:** Beschreibt das Nutzen, welches für den Anwender durch die Resultate erzielt wurde.

4. **Verständlichkeit:** Die Ergebnisse des Modells sollten von einem anderen Anwender verstanden werden.

Anhand dieser Anforderungen sollen die späteren Resultate der modellierten Funktionen gemessen werden. Um dabei eine aussagekräftige Interpretation der Ergebnisse treffen zu können, erfordert es ein hohes Maß an Verständnis der vorliegenden Problemstellung. Dazu bietet sich ein Team von Experten an, welche die Resultate validieren, sodass eine korrekte Bewertung erzielt werden kann. Für die Interpretationsphase eignet sich die Verwendung von Werkzeugen, wie der Visualisierung, um schnellen Aufschluss über die gewonnenen Muster und Zusammenhänge zu erlangen. Innerhalb des iterativen KDD-Prozesses (siehe Abbildung 2 auf S. 9) ist ein Rücksprung in die vorherigen Phasen typisch.[120] Meist müssen Daten nochmal nachbereitet, eine andere Data-Mining-Methode ausgewählt oder sogar Daten neu selektiert werden, wenn das gewünschte Ergebnis sich mit der verwendeten Datenbasis nicht erreichen lässt.[121]

2.3 Funktionsmodellierung

Nachdem die Regressionsanalyse in Kapitel 2.2.4 auf S. 23 als Data-Mining-Methode für diese Arbeit festgelegt wurde, wird der Leser im folgenden Kapitel mit den grundlegenden Bestandteilen der Funktionsmodellierung mit Hilfe der Regression vertraut gemacht (siehe Kapitel 2.3.1 auf S. 28). Dazu werden die unterschiedlichen Modelle der Regression vorgestellt und in Bezug auf die vorliegende Problemstellung bewertet. Anschließend wird in Kapitel 2.3.2 auf S. 28 das Software-Tool *MATrix LABoratory* (*MATLAB*) zur Lösung und graphischen Darstellungen von mathematischen Problemen in Bezug auf die Regressionsanalyse beschrieben, um dessen Konzepte und Funktionsweise für die spätere Umsetzung nachvollziehen zu können.

[120] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 11.
[121] Vgl. ebd.

2.3.1 Regressionsanalyse

2.3.1.1 Methode der kleinste Quadrate

2.3.1.2 Regressionsmodelle

Lineare Regression

Multiple Lineare Regression

Nichtparametrische Regression

2.3.1.3 Bestimmtheitsmaß

2.3.2 MatLab

2.3.2.1 Allgemein

2.3.3 Regressionsanalyse

3 Analysephase

3.1 Expected Goals

3.2 Opta-Spieldaten

test[122] test[123]

[122] Vgl. *Opta-Sports*, F24 Appendices, 2017b, S.1.
[123] Vgl. *Opta-Sports*, The collection process, 2017a, S.1.

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

Glossar

Big Data
Definition folgt → S. 6

Clustering
Definition folgt → S. 17

Cross Industry Standard Process for Data Mining (CRISP-DM)
Definition folgt → S. 7, 9, 10

Data Mining (DM)
Definition folgt → S. 5, 8, 12, 20, 23

Internet of Things (IoT)
Definition folgt → S. 6

K-Nearest Neighbours (KNN)
Definition folgt → S. 16, 20

Knowledge Discovery in Data (KDD)
Definition folgt → S. 7, 8, 10, 11, 22, 24, 26

Künstliche Intelligenz
Definition folgt → S. 4, 23

Machine Learning (ML)
Definition folgt → S. 4, 23

Neuronale Netze (NN)
Definition folgt → S. 23

Online Analytical Processing (OLAP)
Definition folgt → S. 11

Regression
Definition folgt → S. 16, 26

Structured Query Language (SQL)
Definition folgt → S. 11

Super Vector Machine (SVM)
Definition folgt → S. 23

Literaturverzeichnis

Adhikari, Animesh/Adhikari, Jhimli [Advances in Knowledge Discovery in Databases, 2015]: Advances in Knowledge Discovery in Databases. Band 79, Intelligent Systems Reference Library. Cham and s.l.: Springer International Publishing, 2015, ISBN 9783319132112

Aggarwal, Charu C. [Data mining: The textbook, 2015]: Data mining: The textbook. Cham: Springer, 2015, ISBN 978–3–319–14142–8

Anderberg, Michael R. [Cluster Analysis for Applications, 2014]: Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Band 19, Probability and mathematical statistics. Burlington: Elsevier Science, 2014, ISBN 0120576503

Chu, Wesley W. [Data mining and knowledge discovery for big data, 2014]: Data mining and knowledge discovery for big data: Methodologies, challenge and opportunities. Band volume 1, Studies in big data. Heidelberg: Springer, 2014, ISBN 978–3–642–40837–3

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

Fahrmeir, Ludwig et al. [Statistik: Der Weg zur Datenanalyse, 2007]: Statistik: Der Weg zur Datenanalyse. 6. Auflage. Berlin: Springer, 2007, Springer-Lehrbuch, ISBN 978–3–540–69739–8

Fasel, Daniel/Meier, Andreas (Hrsg.) [Big Data: Grundlagen, Systeme und Nutzungspotenziale, 2016]: Big Data: Grundlagen, Systeme und Nutzungspotenziale. Wiesbaden: Springer Vieweg, 2016, Edition HMD, ISBN 9783658115883

Fayyad, Usama/Piatetsky-Shapiro, Gregory/Smyth, Padhraic [From Data Mining to Knowledge Discovery in Databases, 1996]: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17 1996, Nr. 3, 37, ISSN 0738–4602

García, Salvador/Luengo, Julián/Herrera, Francisco [Data preprocessing in data mining, 2015]: Data preprocessing in data mining. Band 72, Intelligent Systems Reference Library. Cham: Springer, 2015, ISBN 978–3–319–10247–4

Han, Jiawei/Kamber, Micheline/Pei, Jian [Data mining: Concepts and techniques, 2012]: Data mining: Concepts and techniques. 3. Auflage. Amsterdam: Elsevier/Morgan Kaufmann, 2012, The Morgan Kaufmann series in data management systems, ISBN 978–0–12–381479–1

Mariscal, Gonzalo/Marbán, Óscar/Fernández, Covadonga [Survey of data mining and knowledge discovery process models, 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

Opta-Sports [The collection process, 2017a]: The collection process. 2017 ⟨URL: `http://www.optasports.com/about/how-we-do-it/the-data-collection-process.aspx`⟩ – Zugriff am 19.01.2017

Opta-Sports [F24 Appendices, 2017b]: F24 Appendices: Elements/attribute/value descriptions. 2017 ⟨URL: `http://www.optasports.com/praxis/documentation/football-feed-appendices/f24-appendices.aspx`⟩ – Zugriff am 19.01.2017

Osei-Bryson, Kweku-Muata/Barclay, Corlane (Hrsg.) [Knowledge discovery process and methods, 2015]: Knowledge discovery process and methods to enhance organizational performance. 2015, ISBN 9781336194304

Runkler, Thomas A. [Data Mining: Modelle und Algorithmen, 2015]: Data Mining: Modelle und Algorithmen intelligenter Datenanalyse. 2. Auflage. Wiesbaden: Springer Vieweg, 2015, Computational Intelligence, ISBN 978–3–8348–2171–3

Shi, Yong et al. [Intelligent knowledge, 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