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

CFT	Curve Fitting Tool
CRISP-DM	Cross Industry Standard Process for Data Mining
DFL	Die Liga – Fußballverband e.V.
DM	Data Mining
GSN	Global Soccer Network
IoT	Internet of Things
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Data
KNN	K-Nearest Neighbours
KPI	Key Perfomance Indicator
MATLAB	MATrix LABoratory
MDKQ	Methode der kleinsten Quadrate
ML	Machine Learning
NN	Neuronale Netze
OLAP	Online Analytical Processing
RSS	Residual Sum of Squares
SAP	eigenständiger Markenname - früher: *Systeme, Anwendungen und Produkte in der Datenverarbeitung*
SQL	Structured Query Language
SVM	Super Vector Machine
TSS	Total Sum of Squares
XML	Extensible Markup 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 wissenschaftlichen Teilfragen:

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 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

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

[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]

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.

Erwartete Ergebnisse:

- Verschiedene Funktionen bei unterschiedlicher Betrachtung:
  - der Auswahl der Daten (Schüsse aus dem Spiel, Standards, . . .)
  - des Winkels zum Tor
  - der Distanz zum Tor
- unterschiedliche Flächen der Funktion im dreidimensionalen Raum:
  - Kegel
  - Teil eines Ellipsoids
  - Spline → Genaue Modellierung der Fläche

Anforderungen		
Nr.	**Beschreibung**	**vorgegeben**
1.	Input-Variablen der Funktion sind die Koordinaten eines Schusses	*ja*
2.	Output-Variabel der Funktion ist die Wahrscheinlichkeit zwischen 0 und 1	*ja*
3.	Es dürfen nur Schüsse berücksichtigt werden, die während des „laufenden“ Spiels abgegeben wurden	*ja*
4.	Eigentore müssen ausgeschlossen werden„laufenden“ Spiels abgegeben wurden	*nein*
5.	Geblockte Schüsse müssen ausgeschlossen werden„laufenden“ Spiels abgegeben wurden	*ja*
6.	Der Ursprung der Funktion (*P(0,0,0)*) liegt in der Mitte der gegnerischen Torlinie	*ja*
7.	Die Funktion muss symmetrisch zu beiden Spielhälften sein (Spiegelung an der gedachten Linie zwischen der generischen und der eigenen Tormitte)	*ja*

Tabelle 1: Auflistung der Anforderungen

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 erhalten (vgl. Kapitel 2.1.1). Darüber hinaus werden die Prozesse des Data Minings (vgl. Kapitel 2.1.2 auf S. 8) beleuchtet, wobei der *Knowledge Discovery in Data* Prozess – methodischer Aufbau der späteren Umsetzung – in Kapitel 2.2 auf S. 13 nochmal ausführlich erläutert.

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 umfassenden Datenmengen zu extrahieren.[3] Es handelt sich um einen Prozess des „*Sammelns, Säuberns, Verarbeitens und Analysierens von Daten, zur Gewinnung von nützlichen Informationen.*“[4] Der weltweit gesammelte Datenbestand erhöht sich immer mehr und stellt Analysten vor die Herausforderung, aus dieser Datenflut wertvolle Informationen und organisiertes Wissen abzuleiten. Erst das heutige „Informationszeitalter“ führte zum Beginn des renommierten Wissenschaftsbereiches 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 maschinelle Lernen (*Machine Learning*[GL] (*ML*)), die Mustererkennung, die Systemtheorie sowie die *Künstliche Intelligenz*[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 verwendeten Terminus des *Gold Minings*, statt den des *Rock or Sand Minings* – da diese Bezeichnung das eigentliche Ziel, die Gewinnung von Wissen, beinhaltet.[12,13]

„Unter Wissen verstehen wir interessante Muster, die allgemein gültig sind, nicht trivial, [sondern]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 wird. Zunächst werden aus den Daten Informationen gewonnen, aus welchen wiederum Wissen abgeleitet werden kann, wobei in diesem Prozess der Wissensextraktion die Datenmenge sukzessive abnimmt und sich verdichtet, wie in Abbildung 1 verdeutlicht.

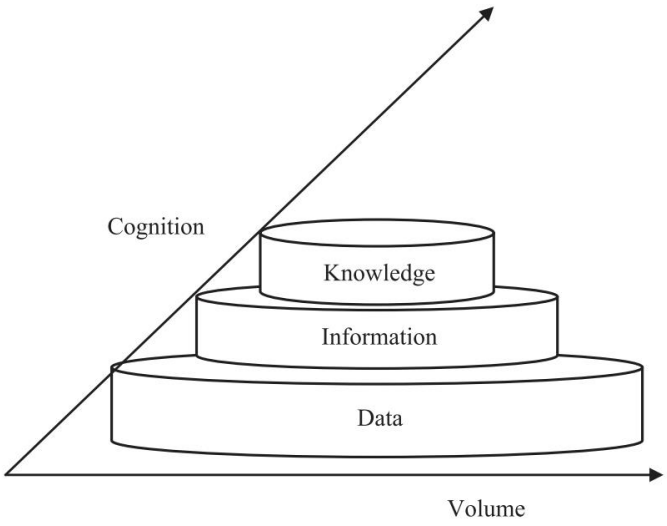

Abbildung 1: Wissensextraktion aus Daten[15]

[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.

Daten stellen dabei nur eine Reihe von Zeichen dar, wobei deren Bedeutung zunächst unklar ist. Erst wenn bekannt wird, in welchem Kontext die Daten stehen und welche Beziehungen zwischen diesen besteht, können diese interpretiert werden und zu einer (relevanten) *Information* heranwachsen. Das *Wissen* entsteht letztlich durch die Verknüpfung von vielen Daten und Informationen sowie den damit gesammelten Erfahrungen.[16]

Durch den Einsatz von modernster Computerhard- als auch Software ist es möglich, sehr große Datenmengen zu erheben, zu verarbeiten und zu analysieren, wodurch in diesem Kontext der Begriff *Big Data*[GL] entstanden ist.[17] Big Data bezeichnet Datenmengen, die mit herkömmlichen Analysemethoden nicht mehr zu verarbeiten sind und deshalb die Anwendung des Data Minings benötigen.[18,19] Hierzu einige ausgewählte Bespiele aus verschiedenen Datenbereichsquellen:[20]

- **World Wide Web:** Die Anzahl der Dokumente im Internet hat seit langem 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 Netzwerksauslastung, 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, Basket-

[16] Vgl. *Shi* et al., Intelligent knowledge, 2015, S. 16-18.
[17] Vgl. *Witten/Frank/Hall*, Data mining: machine learning and techniques, 2011, S. 3.
[18] Vgl. *Fasel/Meier*, Big Data: Grundlagen, Systeme und Nutzungspotenziale, 2016, S. 5.
[19] Vgl. *Shi* et al., Intelligent knowledge, 2015, S. 1.
[20] Vgl. *Aggarwal*, Data mining: The textbook, 2015, S. 2.

ball, Football oder wie in dieser Arbeit Fußball), uvm.[21,22,23]

„Wir befinden uns in einer Welt, in der wir reich an Daten sind, jedoch arm an Informationen und Wissen.“[24] Der unglaublich rapide und enorme Datenzuwachs hat bei Weitem unsere menschliche Vorstellungskraft und Möglichkeiten übertroffen, sodass wir auf effiziente Werkzeuge angewiesen sind (siehe Kapitel 2.2.4 auf S. 24). Die sich immer weiter ausbreitende Lücke zwischen Daten und Informationen, führt nur noch durch die Nutzung von Methoden des Data Minings zu den begehrten „*Golden Nuggets of Knowledge*“.[25] 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.“[26] Das Software-Tool *MatLab* wird innerhalb der Funktionsmodellierung in Kapitel 2.3 auf S. 29 vorgestellt und anschließend als Werkzeug zur Nutzung von Data Mining Methoden in der Umsetzungsphase genutzt (vgl. Kapitel 4 auf S. 49).

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 zueinander.[27,28,29] Schlechte Qualität der Daten mindert die Leistungsfähigkeit des Data Minings. Um die Aussagekraft der Daten nicht zu gefährden, sind vorab Prozessschritte notwendig, die die Daten in adaptierter Form für die Methoden des Data Minings bereitstellen.[30] Hierzu werden im Folgenden kurz die zwei bekanntesten Prozessmodelle vorgestellt:

[21] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 39.
[22] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 1-2.
[23] Vgl. *Chu*, Data mining and knowledge discovery for big data, 2014, S. 85 ff.
[24] *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 5.
[25] Vgl. ebd.
[26] Vgl. *Runkler*, Data Mining: Modelle und Algorithmen, 2015, S. 3.
[27] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 39.
[28] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 2.
[29] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 1.
[30] Vgl. ebd., 2015, S. 10.

- *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 90er-Jahren geprägt und wird als „nicht trivialer Prozess zur Identifizierung von gültigen, neuartigen, potentiell sinnvollen und letztlich verständlichen Muster in Daten“[31] definiert.[32] Erstmals wurde der Terminus von Gregory Piatetsky-Shapiro auf der *International Joint Conference on Artificial Intelligence*, 1989 in Detroit (USA), der Öffentlichkeit präsentiert.[33] Der in Abbildung 2 auf S. 10 dargestellte iterative KDD-Prozess nach Fayyad, beinhaltet folgende Schritte, wobei das DM als ein eigener Prozessschritt ausgewiesen wird:[34]

1. **Datenselektion**: Auswahl der geeigneten Datenmengen.
2. **Datenvorverarbeitung**: Behandlung fehlender oder problembehafteter Daten.
3. **Datentransformation**: Umwandlung in adäquate Datenformate.
4. **Data Mining**: Suche nach Mustern.
5. **Interpretation und Evaluation**: Interpretation der Ergebnisse und Auswertung derer.

Auf die einzelnen Prozessschritte und deren Methoden wird genauer in Kapitel 2.2 auf S. 13 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*.[35] Alle zielen dabei auf die Erforschung von Wissen aus Datenmengen ab, wodurch in dieser

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

Arbeit die allgemeingültige Bezeichnung *Knowledge Discovery in Data* verwendet wird.

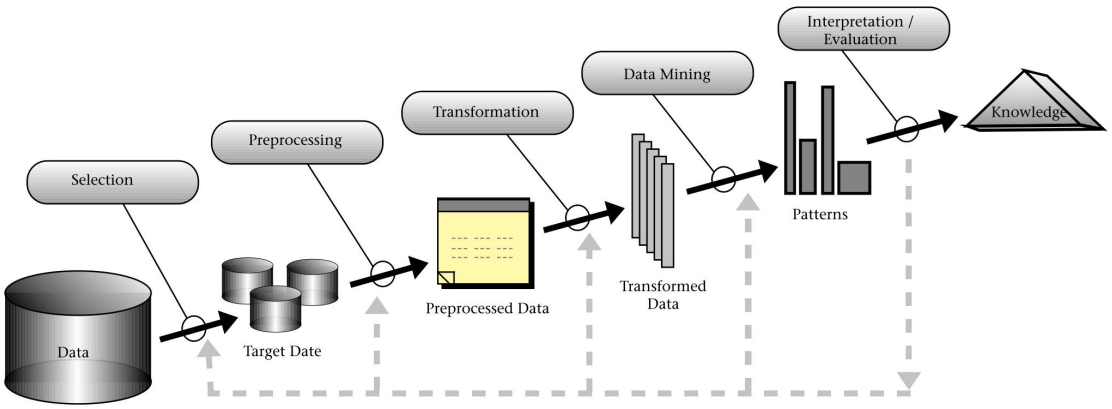

Abbildung 2: Der Knowledge Discovery in Data Prozess[36]

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:[37,38]

- 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.

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

Hierbei wird von einem Lebenszyklus mit sechs beinhaltenden Etappen ausgegangen, die in Abbildung 3 dargestellt werden.[39] Im Folgenden werden dazu die einzelnen Schritte des Prozesses aus der Abbildung (nummeriert von Schritt 1 bis 6) kurz beschrieben.

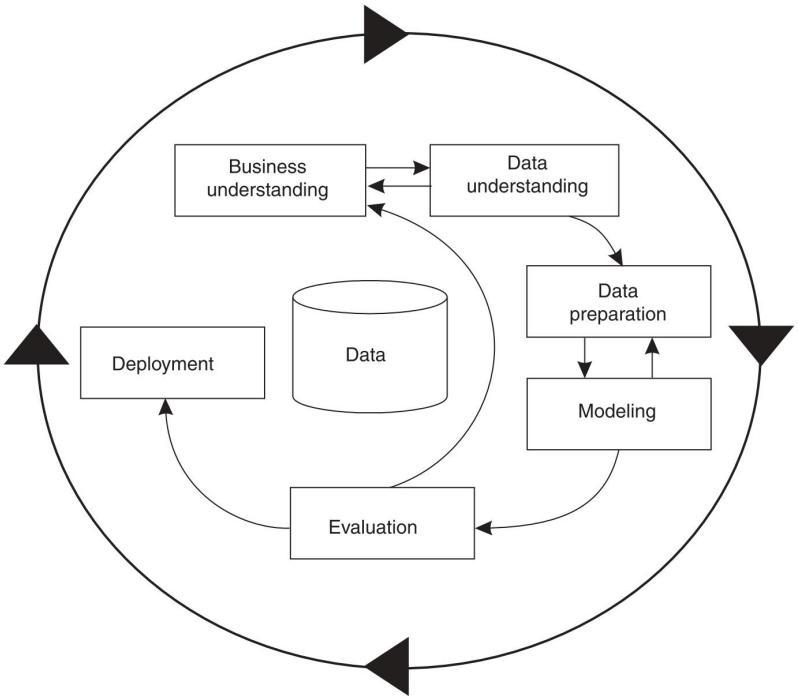

Abbildung 3: CRISP-DM Prozess[40]

1. Verstehen der Aufgabe (*Business understanding*): 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 (*Data understanding*): Diese Phase beschäftigt sich mit den benötigten Daten zur Durchführung der Analyse. Daten werden gesammelt und beschrieben, um deren betriebliche Bedeutung zu verstehen.

[39] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 6-8.
[40] Vgl. Abbildung *Mariscal* et al., A survey of data mining, 2010, S. 13

3. Datenvorbereitung (*Data preparation*): 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 (*Modeling*): In dieser Phase wird ein Modell mit Hilfe des Data Minings erstellt, welches durch einen iterativen Aufbau immer wieder verfeinert und verbessert wird.

5. Evaluation: 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 (*Deployment*): Zuletzt gilt es den Einsatz der Resultate in dem Unternehmen vorzubereiten und diese 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.[41]

Im Vergleich zum KDD-Modell nach Fayyad, sind die Phasen 1 und 2 des CRISP-DM-Modells sehr stark projektabhängig und spiegeln die Sicht der Industrie auf das Projekt wider.[42] 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 auf S. 13 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.[43] Nicht zuletzt ist zu

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

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 wissenschaftlichen 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. 9) 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 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 sollte. Eine einfache *Structured Query Language*^{GL} (*SQL*)-Abfrage oder *Online Analytical Processing*^{GL} (*OLAP*) reichen für eine komplexe Datenanalyse oft nicht aus.“[44] Hier greift der in Abbildung 2 auf S. 10 dargestellte KDD-Prozess, ein multiples iteratives Modell, in dem die einzelnen Schritte solange wiederholt und aufeinander abgestimmt werden müssen, bis aus den zugrundeliegenden Daten, Wissen abgeleitet werden kann.[45] Das Data Mining selbst kommt erst nach ausführlicher Datenvorbereitung zum Einsatz und kann so zu einer automatischen und explorativen Anpassung eines Modells – wie bei der Funktionsmodellierung (vgl. Kapitel 2.3 auf S. 29) – an riesige Datenmengen genutzt werden.[46,47]

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.[48,49] Im Folgenden wird

[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. *Adhikari/Adhikari*, Advances in Knowledge Discovery in Databases, 2015, S. 1.
[47] Vgl. *Mariscal/Marbán/Fernández*, Survey of data mining and knowledge discovery process models, 2010, S. 7.
[48] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 1.
[49] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 198.

versucht, diese Schritte klar voneinander 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 stattfindet.[50] Der Datenanalyst befasst sich in dieser Phase mit der Bestimmung der für die Analyse geeigneten Daten und dem Export dieser Datenauswahl beispielsweise in eine Datenbank. Die selektierten Daten können zum Beispiel technischen oder rechtlichen Restriktionen unterliegen, wie zum Beispiel Zugriffs- oder Kapazitätsbeschränkungen. Hierbei sollte auf eine repräsentative Teilmenge des Datenbestandes zurückgegriffen werden.[51]

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.“[52]

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 in einem 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.[53] Es gilt auch hier das Prinzip: GIGO – garbage in, garbage out.[54] Die oftmals schlechte Qualität der (Roh-)Daten ist durch *fehlende, ungenaue, inkonsistente bzw. widersprüchliche* Daten zu begründen.[55,56] Im Folgenden werden dazu einige Ursachen beispielhaft aufgeführt.

[50] Vgl. *Fayyad/Piatetsky-Shapiro/Smyth*, From Data Mining to Knowledge Discovery in Databases, 1996, S. 42.
[51] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 9.
[52] Ebd.
[53] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 10-11.
[54] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 197.
[55] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 84.
[56] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 196.

Ungenaue bzw. falsche Daten können schon bei der Erhebung entstehen, wenn ein falsches Datenerhebungsinstrument ausgewählt wird. Bei Stichproben sollte die Gesamtmenge so präzise wie möglich widergespiegelt werden, um die Datenakkuratesse nicht zu gefährden.[57] 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.[58,59,60] Neben der falschen subjektiven Einschätzung des Menschen bei der Erhebung, können auch von einem technischem Blickwinkel ungenaue Daten ermittelt werden, wie z.B. durch (teils-)defekte Sensoren. Nicht zuletzt können Daten bei einem Transfer verfälscht werden bzw. sogar verloren gehen.[61]

Fehlende Daten lassen sich einerseits durch technische Mängel begründen, andererseits auch durch die Tatsache, dass bestimmte Attribute schlichtweg von Beginn an bei der Erhebung nicht beachtet wurden oder durch bestimmte Restriktionen nicht verfügbar waren.[62]

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 leistungsstarke Werkzeuge zur Verfügung, um die Datenqualität nachhaltig zu verbessern:[63,64,65]

- **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. 16).

- **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. 19).

- **Data Reduction**: Um die Algorithmen der Data Mining Methoden nutzen zu können, muss die exorbitante Datenmenge reduziert bzw. komprimiert werden, sodass lange Laufzeiten verhindert beziehungsweise reduziert werden können (siehe Kapitel 2.2.2.3 auf S. 21).

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

Auf die in Abbildung 4 vereinfacht dargestellten Werkzeuge und ihre Konzepte, wird in den folgenden Unterkapiteln näher eingegangen.

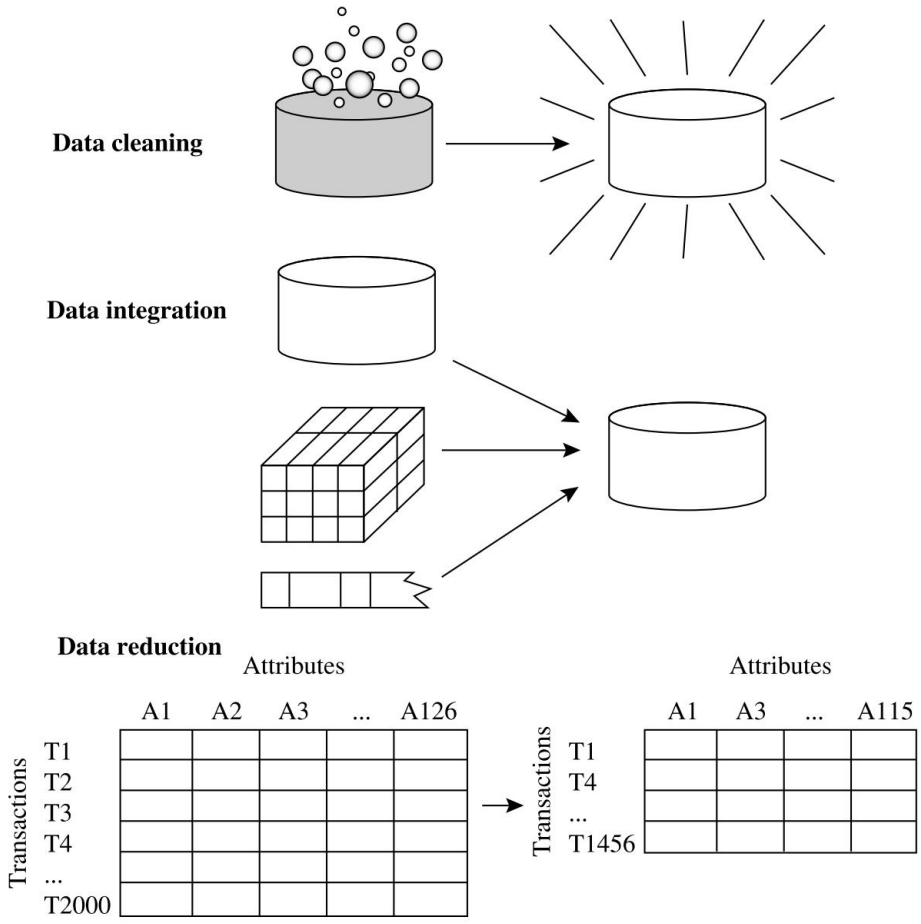

Abbildung 4: Werkzeuge der Datenvorverarbeitung[66]

2.2.2.1 Data Cleaning

In der realen Welt sind Daten häufig „unvollständig, mit Fehlern oder Ausreißern behaftet oder sogar inkonsistent.“[67] 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 bedeutet, es sollen möglichst keine neuen Informationen hinzugefügt werden,

[66] Vgl. Abbildung *Han*, Data Mining: Concepts and techniques, 2012, S. 87
[67] *Cleve/Lämmel*, Data Mining, 2014, S. 199.

die das reale Abbild verzerren oder verfälschen könnten.[68] Folgende Problemarten gilt es zu behandeln:

Fehlende Daten Dem Datenanalyst stehen einige Möglichkeiten zur Verfügung, um auf fehlende Daten zu reagieren:[69,70]

- *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 schwer zu realisierenden 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 die Berechnung nur auf Datensätzen der selben Klasse angewandt wird. Die Methode der *K-Nearest Neighbours*$^{\text{GL}}$ (*KNN*)[71] steht zur Verfügung, wenn keine Klassen vorhanden sind. Hierbei wird der Durchschnitt, der dem aktuellen Datensatz ähnlichsten Werte benutzt.

- *Wahrscheinlichster oder häufigster Wert*
  Durch statistische Methoden kann der wahrscheinlichste Wert für das fehlende Attribut ermittelt werden, jedoch sollte diese Angleichung begründet sein. Bei

[68] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 199-200.
[69] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 88-90.
[70] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 200-201.
[71] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 76.

nicht numerischen Werten kann als weitere Möglichkeit auch der häufigste Wert, als Ersatz für das fehlende Attribut verwendet werden.

Verrauschte Daten und Ausreißer Durch ungenaue Messwerte oder falschen Schätzungen entstehen die sogenannten *verrauschten Daten*.[72] Um diese bereinigen zu können, stehen dem Datenanalyst einige Verfahren zur Verfügung, wodurch diese fehlerbehafteten Daten angeglichen werden können.[73] Als *Ausreißer* bezeichnet man dabei Daten, die erheblich von den anderen Daten abweichen oder außerhalb eines Wertebereiches liegen.[74] 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.[75] „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.“[76]

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

- *Regression*[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. 29).

- *Verbundbildung (clustering)*
  Eine der einfachsten Möglichkeiten um Ausreißer zu erkennen, bietet die Verbundbildung, auch *Clustering*[GL] genannt. Hierbei werden ähnliche Daten, wie in Abbildung 5 auf S. 19 dargestellt, zu *Clustern* zusammengeführt, wodurch sich die Ausreißer direkt identifizieren lassen.

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

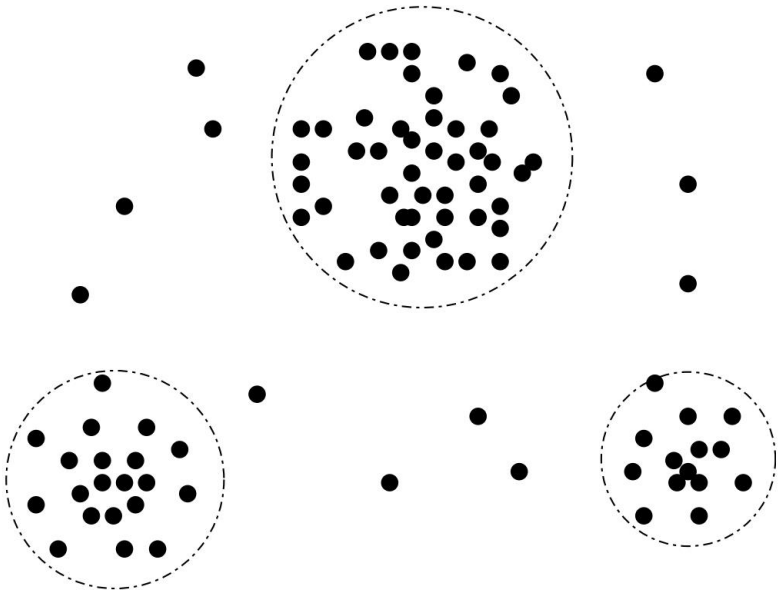

Abbildung 5: Outlierdetection mittels Clustering[77]

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.[78]

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.[79] Folgende Punkte gilt es bei der Daten-

[77] Vgl. Abbildung *Han*, Data Mining: Concepts and techniques, 2012, S. 91
[78] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 203-204.
[79] 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.[80,81] 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.[82]

- **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.[83] 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[84] bestimmen zu können, verwendet man in der Regel den χ^2(*Chi*^{2})-Test.[85,86,87]

- **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

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

Analyseergebnissen. Ein häufiger Grund stellt dabei die Verwendung von denormalisierten Datenbanktabellen dar.[88,89]

- **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.[90,91]

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:[92,93,94]

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.[95]

[88] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 98.
[89] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 43.
[90] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 99.
[91] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 199.
[92] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 147 ff.
[93] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 99-100.
[94] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 206-208.
[95] 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.[96]

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.[97,98]

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.[99] Oftmals müssen sogar neue Attribute aus einem Datensatz kreiert werden, da diese nicht in geeigneter Struktur für das Data-Mining-Verfahren vorliegen.[100] 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:[101]

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

[96] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 207.
[97] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 25-27.
[98] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 207.
[99] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 112.
[100] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 48.
[101] 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.[102]

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.[103,104]

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

Abbildung 6: Min-Max-Normalisierung

Datenaggregation Nicht nur aus Sicht der Datenkompression (vgl. Kapitel 2.2.2.3 auf S. 22) ist die Datenaggregation erforderlich. Vielmehr „kann die Aggregation aus inhaltlichen Gründen sinnvoll sein."[105] 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.[106]

Datenglättung Die bereits in Kapitel 2.2.2.1 auf S. 18 vorgestellten Techniken zur Bereinigung von verrauschten Daten und Ausreißer, finden auch bei der Transformation ihre Verwendung. Die Datenglättung strebt nach einer reduzierten Datenmenge,

[102] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 211.
[103] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 114.
[104] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 212.
[105] Ebd., 2014, S. 214.
[106] 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.[107]

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.[108] 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.[109] Im allgemeinen unterscheidet man die Data-Mining-Methoden in zweit Kategorien: *Prognose* und *Beschreibung*. Hierzu gibt Abbildung 7 auf S. 27 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.[110] Die am häufigst angewendeten statistischen Methoden sind:[111,112]

- *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.[113]

[107] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 214-215.
[108] Vgl. ebd., 2014, S. 10.
[109] Vgl. ebd., 2014, S. 12.
[110] Vgl. *García/Luengo/Herrera*, Data preprocessing in data mining, 2015, S. 3.
[111] Vgl. ebd., 2015, S. 3-5.
[112] Vgl. *Han/Kamber/Pei*, Data mining: Concepts and techniques, 2012, S. 23-24.
[113] 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.[114]

- *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.[115]

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.[116]

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*).[117]

- *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 werden.

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

Visualisierung als Werkzeug: 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.[118] 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.[119]

Auswahl der Methode: 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. 27 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. 29 betrachtet, um diese Technik in der späteren Umsetzung anwenden zu können.

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

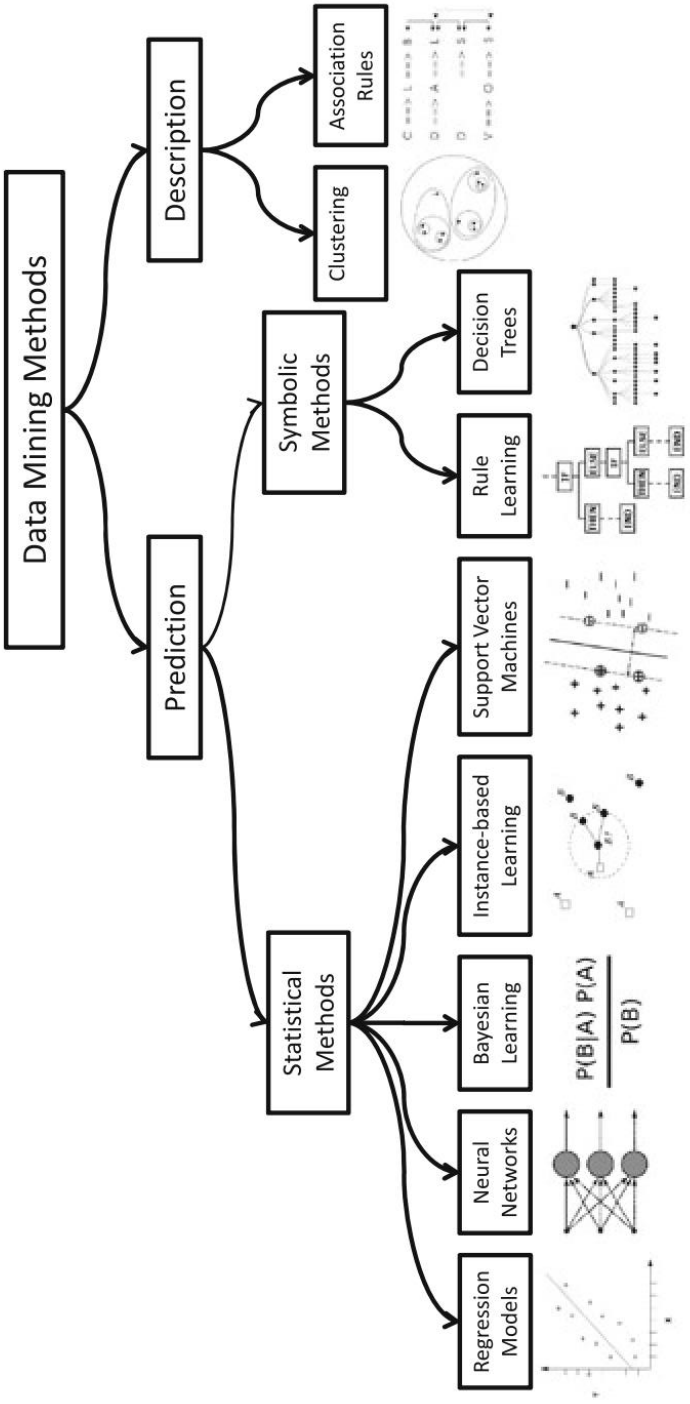

Abbildung 7: Übersicht: Data-Mining-Methoden[120]

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

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:[121]

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 erworbene 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. 10) ist ein Rücksprung in die vorherigen Phasen typisch.[122] 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.[123]

[121] Vgl. *Cleve/Lämmel*, Data Mining, 2014, S. 11-12.
[122] Vgl. ebd., 2014, S. 11.
[123] Vgl. ebd.

2.3 Funktionsmodellierung

Nachdem die Regressionsanalyse in Kapitel 2.2.4 auf S. 24 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). 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. 39 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.

2.3.1 Regressionsanalyse

2.3.1.1 Allgemein

Die Regression (lat. *regredi* für umkehren, zurückkehren) beinhaltet im Allgemeinen die Analyse einer abhängigen Variablen von einer oder mehreren unabhängigen Variablen.[124] Dabei drücken die unabhängigen Variablen die abhängige Variable mittels einer *Regressionsgleichung* aus.[125] Die in der mathematischen Gleichung beinhaltenden Parameter (auch *Regressoren* genannt) müssen so gewählt und justiert werden, dass eine möglichst genau Anpassung der resultierenden Funktion an die vorhandenen Daten erzielt wird.[126,127] Diese rein datenbasierte mathematische Beschreibung hat ihren Ursprung in einer Studie von Francis Galton[128], in der die Körpergröße von Kindern in Bezug zu derer ihrer Eltern analysiert wurde.[129] Anhand der vorliegenden Problemstellung dieser Arbeit, lässt sich das Vorgehen der Regressionsanalyse beispielhaft demonstrieren:

Es wird versucht, die Abhängigkeit der Wahrscheinlichkeit eines Torerfolges (=*abhängige Variable*), von den Einflussfaktoren, wie der Distanz oder des Winkel zum Tor bzw. der Koordinaten des Schusses (=*unabhängige Variablen*), mit Hilfe einer Funktion (=*Regressionsgleichung*) zu ermitteln.

[124] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 5.
[125] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 475.
[126] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 68.
[127] Vgl. *Schimek*, Smoothing and regression, 2000, S. 1-2.
[128] britischer Naturforscher im 19. Jahrhundert - prägte erstmals den Begriff der *Regression*
[129] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 68.

Anhand der *linearen Regression* sollen die grundlegenden Bestandteile aller Regressionsmodelle erläutert werden. Um eine Punktewolke (die beobachteten Daten) durch eine Funktion $\hat{f}$ zu approximieren, bedient man sich der quadratischen Abstände der Datenpunkte zur Funktion und versucht diese durch die *Methode der kleinsten Quadrate*$^{\text{GL}}$ (*MDKQ*) zu minimieren.[130] Die lineare Regressionsfunktion dabei liegt in der Form

$$\hat{f}(x) = \hat{\alpha} \cdot x + \hat{\beta} \tag{2}$$

vor.[131] Wie in Abbildung 8 auf S. 31 exemplarisch dargestellt ist, werden die Abstände zwischen den beobachteten Datenpunkten (orangefarbene Kreise) und der Funktion $(y_i - \hat{y}(x_i) \rightarrow i = 1, ..., m)$ summiert, woraus sich die *Summe der Fehlerquadrate* (=*Residual Sum of Squares* (*RSS*)) ergibt:[132]

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 \tag{3}$$

Durch die Berechnung der RSS wird die Distanz zwischen den Daten und dem Modell berechnet. Um eine möglichst genaue Abbildung der Daten durch das Modell gewährleisten zu können, müssen die Parameter $\hat{\alpha}$ und $\hat{\beta}$ so gewählt werden, dass die Summe der Fehlerquadrate minimal ist.[133] Es gilt:

$$\min_{a,b \in \mathbb{R}} RSS \tag{4}$$

Das mathematische Vorgehen für die Ermittlung eines Minimums einer Funktion mit mehreren Variablen – partielle Ableitung von $\text{RSS}(\hat{\alpha}, \hat{\beta})$ – wird verwendet, um die Parameter der Regressionsfunktion ausfindig zu machen:[134]

$$\hat{\alpha} = \frac{\sum_{i=1}^{m} x_i y_i - m\overline{xy}}{\sum_{i=1}^{m} x_i^2 - m\overline{x}^2} \tag{5}$$

[130] Vgl. *Hastie/ Tibshirani/ Friedman*, The elements of statistical learning, 2016, S. 44.
[131] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 476.
[132] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 37.
[133] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 69.
[134] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 39.

$$\hat{\beta} = \overline{y} - \hat{\alpha}\overline{x}^2 \tag{6}$$

Diese grundlegenden Bestandteile finden sich in allen Regressionsmodellen wieder, die in Kapitel 2.3.1.2 kurz vorgestellt werden.

2.3.1.2 Regressionsmodelle

In der Praxis haben sich durch den allgemein Ansatz der Regression eine Vielzahl von Modellen etabliert, die je nach Anwendungsfall ihre Verwendungen finden:

- **Lineare Regression**
  Am bekanntesten ist die lineare Regression, die auch oft als „Ausgleichsgerade“ bezeichnet wird und zur Prognose einer unabhängigen Größe y, in Abhängigkeit *einer* bekannten Größe x, angewendet wird.[135] Eine Funktion $\hat{f}(x)$ wie Gleichung 2 auf S. 30 ist von ihren Regressoren α und β *linear* abhängig und heißt deshalb *lineare Regressionsfunktion*.[136] In Abbildung 8 ist eine solche Funktion beispielhaft dargestellt.

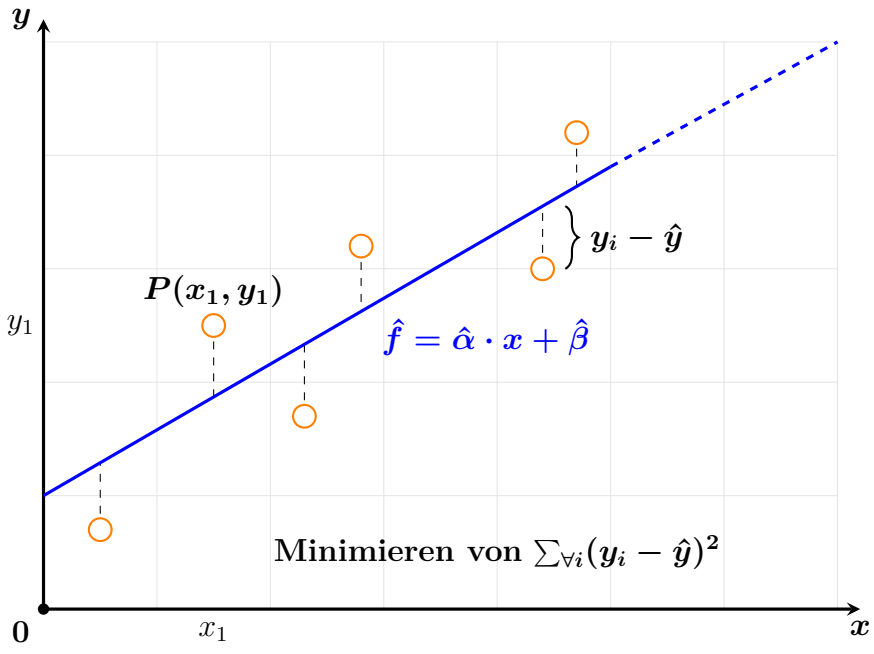

Abbildung 8: Grafische Darstellung der linearen Regression[137]

[135] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 475.
[136] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 68.

- **Nichtlineare Regression**

  In vielen realen Anwendungen kann die Regressionsfunktion nicht durch die Linearkombination der Regressionskoeffizienten berechnet werden, da diese in *linearer* Weise von den Regressoren abhängt. Allgemein lässt sich dieses Modell mit $\boldsymbol{x} = (x_1, ..., x_n)$ und $\boldsymbol{a} = (a_1, ..., x_s)$ wie folgt ausdrücken:[138]

$$\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{a}) \tag{7}$$

  In Abbildung 9 auf S. 33 liegen die Daten in einem oszillierenden, sinusförmigen Muster vor und können folglich mit der allgemeinen Sinusfunktion beschrieben werden ($\hat{f} = \alpha_0 \cdot \sin(\alpha_1 \cdot (x - \alpha_2))$). Bei dieser nichtlinearen Regressionsfunktionen können die Regressoren dazu verwendet werden, die Funktion möglichst genau an die Daten anzupassen, wobei α_0 die Amplitude bestimmt, α_1 die Periode und α_2 die Sinus-Funktion entlang der x-Achse verschiebt.[139] Auch hier lassen sich mit Hilfe des Standardmodells, der Minimierung der kleinsten Quadrate, die Parameter bestimmen.[140]

[137] Abbildung in Anlehnung an *Backhaus* et al., Multivariate Analysemethoden, 2016, S. 71.
[138] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 85.
[139] Vgl. ebd., 2014, S. 85-86.
[140] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 509.

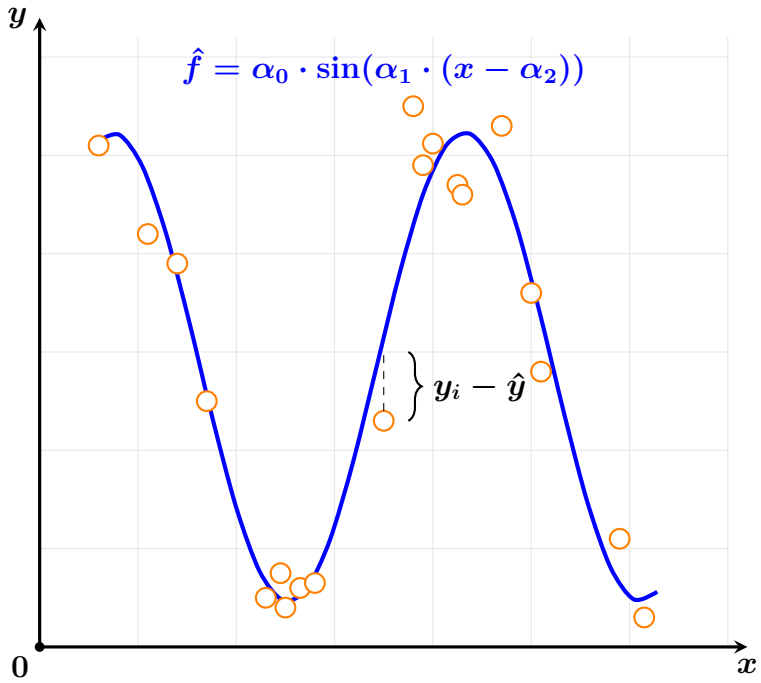

Abbildung 9: Grafische Darstellung der nichtlinearen Regression[141]

- **Multiple Regression** In den meisten technischen und wirtschaftlichen Anwendungsfällen ist die Zielvariable y von mehr als einer unabhängigen Variable abhängig. Dieser Fall kann durch die *multiple* oder auch *multivariate Regression* behandelt werden, wobei sich diese Regressionsfunktion mit den unabhängigen Variablen ($\boldsymbol{x} = (x_1, ..., x_n)$) und beliebigen reellen Funktionen (f_i) im Allgemeinen wie folgt ausdrücken lässt:[142]

$$\hat{f}(\boldsymbol{x}) = \alpha_0 + \alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_1(\boldsymbol{x}) + ... + \alpha_s f_s(\boldsymbol{x}) \tag{8}$$

  Die Vorgehensweise zur Ermittlung der Parameter entspricht der, der linearen Regression – auch hier wird die Methode der Minimierung der kleinsten Quadrate angewendet. Liegt beispielsweise eine Punktewolke wie in Abbildung 10 auf S. 34 vor, kann diese durch die Funktion

$$\hat{f}(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \tag{9}$$

[141] Abbildung in Anlehnung an *Günther* et al., Mathematische Modellbildung und Simulation, 2014, S. 86.
[142] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 41-42.

beschrieben werden, wobei der Regressor α_0 die Verschiebung der Fläche entlang der y-Achse, α_1 die Steigung der Variable x_1 sowie α_2 die Steigung von x_2 angibt.

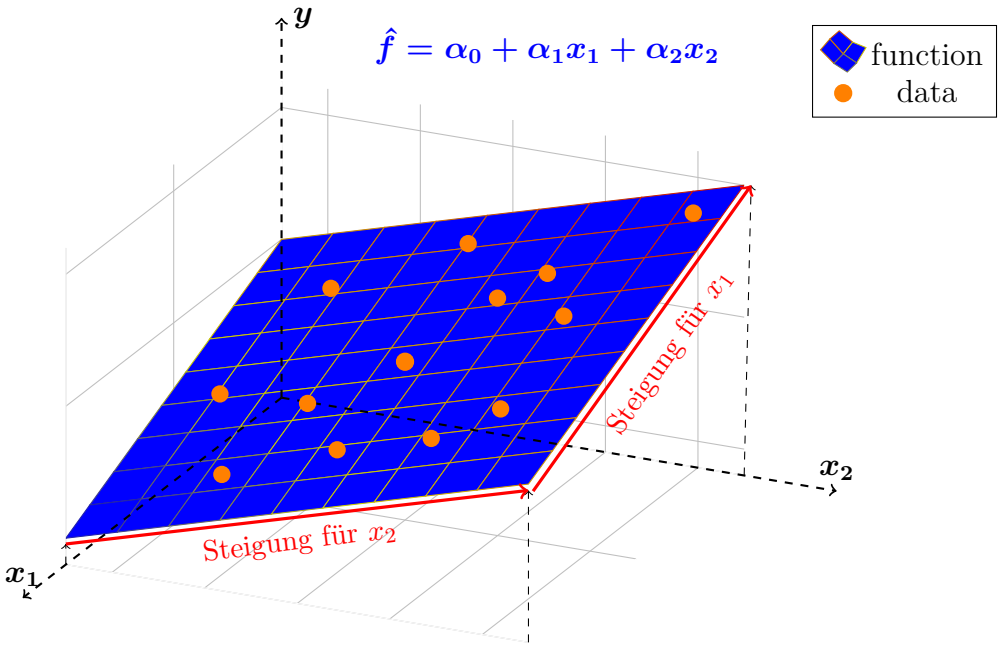

Abbildung 10: Grafische Darstellung der multiplen Regression[143]

- **Nichtparametrische Regression** In vielen Anwendungen lässt sich nicht von vornherein eine Regressionsfunktion mit bestimmter Spezifizierung der Parameter vorhersagen. In den vorangehenden Beispielen – ob linear oder nichtlinear – wurde jeweils ein konkreter Ausdruck vorgegeben, um mittels Minimierung die Funktion an die Daten anzupassen. Betrachtet man Abbildung 11 auf S. 35, so lässt sich schnell erkennen, dass es dazu keine passende mathematische Funktion geben wird.[144] Die nichtparametrischen Regression verfolgt das Ziel, die Funktion $\hat{f}$ möglichst genau zu schätzen. Etabliert haben sich hierbei Methoden wie *Spline-Regressionen* und *lokale Regressionsschätzer*, die jedoch aufgrund ihres numerischen Aufwands hier nicht detailliert beschrieben werden und selbst nur durch bei der Verwendung von statistischen Programmpaketen (vgl. MATLAB Kapitel 2.3.2 auf S. 39) Anwendungen finden.[145]

[143] Eigene Darstellung
[144] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S. 91.
[145] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 510.

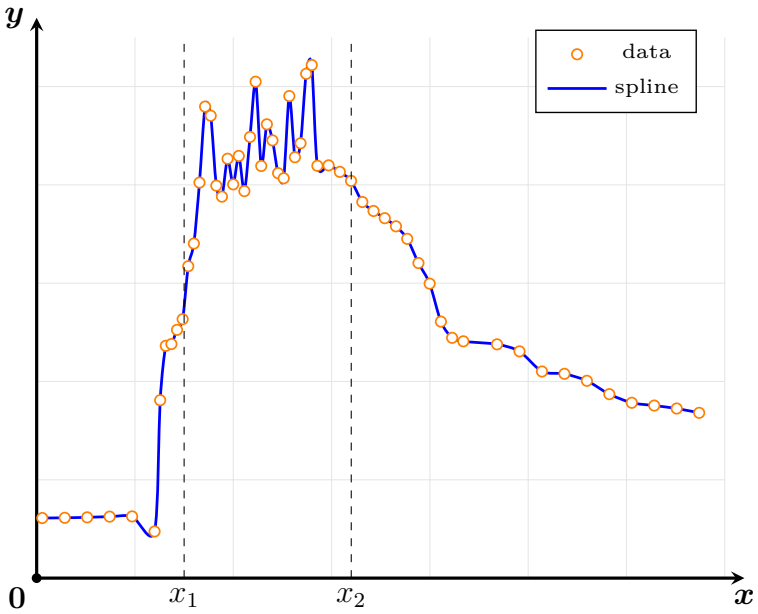

Abbildung 11: Grafische Darstellung der nichtparametrischen Regression[146]

„Splinefunktionen gehören zu den wichtigsten und verbreitetsten Regressionsmethoden und werden quer durch alle Disziplinen z.B. in Betriebswirtschaft, Informatik, Bildverarbeitung, Medizin, Maschinen.“[147]

Der Gedanke der *smoothing splines*[148] ist, den Bereich der x-Werte durch ein feines Gitter so zu unterteilen, das sich die angrenzenden Intervalle durch glatt miteinander verbundene Polynomfunktionen niedrigen Grades (oftmals kubisches Polynom) approximieren lassen.[149] In Abbildung 12 auf S. 36 ist dazu die Punktewolke aus Abbildung 11 im Wertebereich zwischen x_1 und x_2 genauer dargestellt, um das Resultat dieses Verfahren besser betrachten zu können. Legt man das Augenmerk ausschließlich auf den Intervallbereich I_i, so lässt sich dieser Bereich durch ein kubisches Polynom ausdrücken. Werden all diese Intervalle als eigene Funktionen definiert und stückweise an den „Knotenpunkten“ (Übergang zwischen den einzelnen Intervallen) stetig und differenziert aneinander gesetzt, erhält man die gesuchte Regressionsfunktion $\hat{f}(x)$.[150]

[146] Abbildung in Anlehnung an *Günther* et al., Mathematische Modellbildung und Simulation, 2014, S. 92.
[147] *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S.93.
[148] Die englische Übersetzung bedeutet so viel wie *glättende Verzahnung*
[149] Vgl. *Fahrmeir* et al., Statistik: Der Weg zur Datenanalyse, 2007, S. 510.
[150] Vgl. ebd.

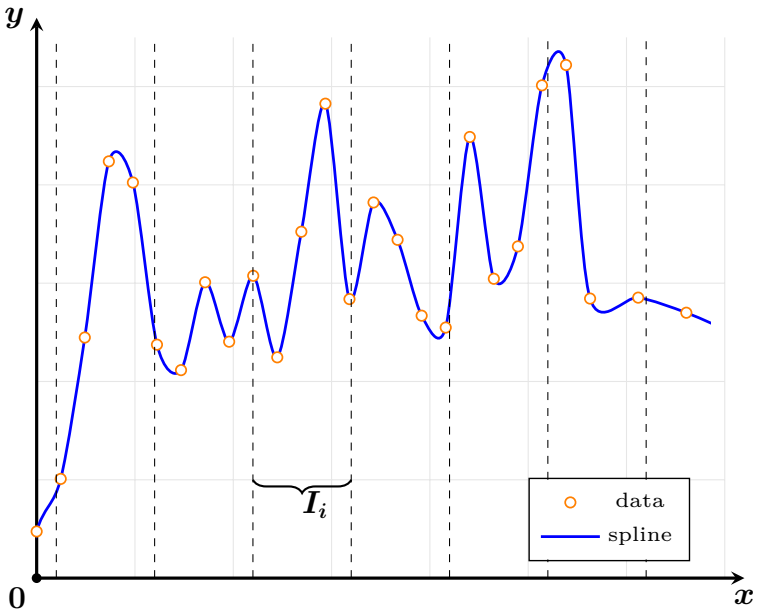

Abbildung 12: Anwendung der *Spline*-Regression[151]

Bewertung Wie in der Einleitung dieses Kapitels beschrieben, wird die Wahrscheinlichkeit eines Torerfolges durch mehrere Faktoren, wie beispielsweise den Koordinaten des Schusses, beeinflusst, wodurch die *lineare Regression* für die Modellierung ausgeschlossen werden kann, da die Zielvariable in der vorliegenden Problemstellung von mindestens zwei unabhängigen Variablen abhängt. In einem multiplen Regressionsmodell können mehrere unabhängige Variablen behandelt werden, jedoch müsste auch hier von vornherein eine parametrische Spezifikation der Funktion angegeben werden. Eine erste mögliche Vorstellung in Bezug wäre dazu die Funktion $\hat{f}(x_1, x_2) = \alpha_0 e^{-x_1} \cdot \alpha_1 \sin(\alpha_2 \cdot (\pi x_2 - \alpha_3))$, die in Abbildung 13 auf S. 37 abgebildet ist.

[151] Eigene Darstellung: Vergrößerung der Punktewolke aus Abbildung 11 auf S. 35 im Wertebereich x_1 bis x_2

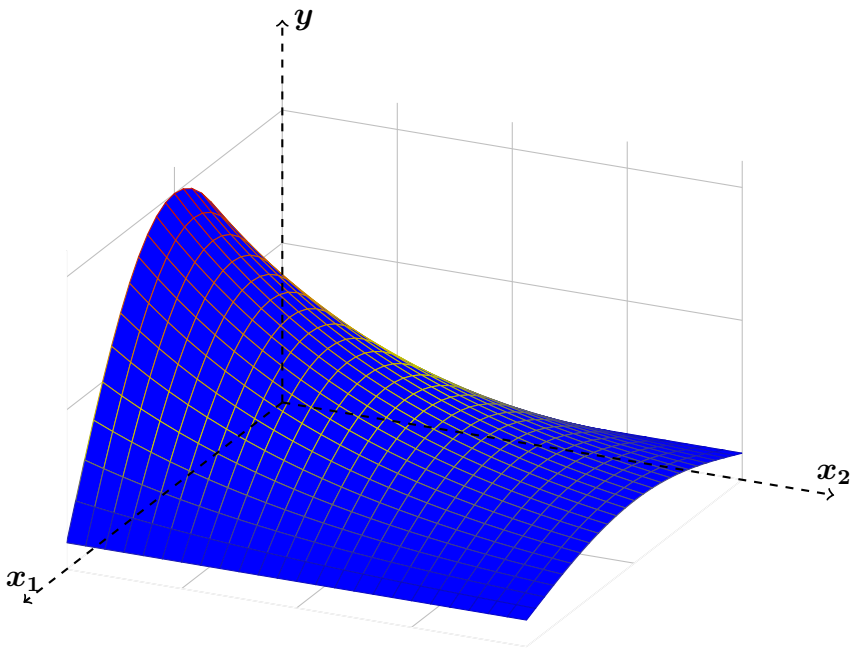

Abbildung 13: Erste Vorstellung der Funktion in Bezug auf die Koordinaten des Schusses[152]

Die Parameter x_1 (=Breite des Spielfeldes) und x_2 (=Länge des Spielfeldes) geben hierbei die Koordinaten den Punkt an, von dem aus geschossen wurde, wobei das gegnerische Tor auf der x_1-Achse – also bei $x_2 = 0$ – mittig platziert ist. Die Vermutung ist, dass sich die Wahrscheinlichkeit y mit zunehmender Nähe zum Tor steigt, wodurch die Fläche in Richtung Tor stetig erhöht und eine Art „Gipfel“ entsteht. Bereits bei der Betrachtung dieser Vermutung, ist zu erkennen, das eine „herkömmliche“ mathematische Funktion das Ergebnis zu einer sehr verfälschenden Glättung führt. Wenn beispielsweise ein Schuss auf der Höhe der Grundlinie seitlich des Tores abgegeben wurde, dann ist allein aufgrund des Winkels ein Torerfolg fast unmöglich und die Wahrscheinlichkeit somit falsch repräsentiert. Folglich ist es sinnvoll, *nichtparametrische Regressionsfunktionen* in Form von *Splines* zu verwenden, um eine exakte Anpassung der Funktion an die vorliegenden Daten zu erreichen. Eine detaillierte Gegenüberstellung der multiplen und nichtparametrischen Regression wird innerhalb der Umsetzung in Kapitel 4.4 auf S. 56 aufgezeigt.

[152] Eigene Darstellung

2.3.1.3 Bestimmtheitsmaß

Um die Qualität der Anpassung des Modells an die Daten zu überprüfen, stellt der graphische Vergleich zwischen Modell und Daten die simpelste Möglichkeit dar. Betrachtet man nochmals die Sinus-Funktion aus Abbildung 9 auf S. 33, kann schnell festgestellt werden, dass das Modell sehr gut an die Daten angepasst wurde.[153] Für eine exakte Modellierung wird ist jedoch Bestimmtheitsmaß R^2 notwendig, welches den *goodness of fit* (*dt. Anpassungsgüte*) misst.[154] Die Qualität des Modells wird dabei auf einer Skala zwischen 0 und 1 dargestellt, wobei ein sehr hoher Wert für eine gute Anpassung und ein niedriger Wert für eine schlechte Anpassung des Modells an die Daten spricht. Diese Skala dient dazu, um verschiedene Regressionsmodelle miteinander vergleichen zu können.[155] Das Bestimmtheitsmaß R^2 misst dabei zu welchem prozentualen Anteil die Abweichung der gemessenen abhängigen Variablen durch die unabhängigen Variablen des Modells erklärt wird und ist formal wie folgt definiert:[156,157,158]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{RSS}{TSS} \tag{10}$$

Wie in Gleichung 10 zu erkennen ist, kann das R^2 auch als Verhältnis von RSS zu *Total Sum of Squares* (*TSS*), also der erklärten Variation zur gesamten Abweichungsqudratsumme, dargestellt werden.[159,160] Die Hinzunahme von weiteren erklärenden x-Variablen führt im schlechtesten Fall dazu, dass das Bestimmtheitsmaß gleich bleibt, unabhängig davon ob die zusätzlichen Variablen die Qualität des Modells verbessern. Man spricht dabei von einer *Überparametrisierung* des Modells.[161] Um diesen Fall zu vermeiden, verwendet man in der Praxis das *korrigierte Bestimmtheitsmaß* $\overline{R}^2$ (*engl.: adjusted* R^2), dass die Hinzunahme von Variablen mit geringer Erklärungskraft „bestraft“ (die Qualität der Anpassung sinkt).[162] Die Aufnahme einer zusätzlichen Variable sollte nur dann in Erwägung gezogen werden, wenn der

[153] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S.71.
[154] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 51.
[155] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S.71.
[156] Vgl. *Daróczi*, Mastering data analysis, 2015, S.118.
[157] Vgl. *Günther/ Velten*, Mathematische Modellbildung und Simulation, 2014, S.72.
[158] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 51.
[159] Vgl. ebd., 2014, S. 48.
[160] Vgl. *Daróczi*, Mastering data analysis, 2015, S.119.
[161] Vgl. *Cleff*, Deskriptive Statistik und moderne Datenanalyse, 2008, S. 160.
[162] Vgl. ebd., 2008, S. 161.

dadurch gewonnene Erklärungswert für das Modell größer als der „Bestrafungsabschlag“ des korrigierten Bestimmtheitsmaßes ist. Das $\overline{R}^2$ kann daher zum Vergleich von Regressionsmodellen mit unterschiedlicher Anzahl von unabhängigen Variablen herangezogen werden, um die Anpassungsgüte des Modells an die Daten zu messen.[163] Die ursprüngliche Interpretation von R^2 geht jedoch durch Bestrafung der Hinzunahme weiterer Parameter weites gehend verloren, sodass beide Bestimmtsheitsmaße für eine Bewertung herangezogen werden sollten.[164]

In diesem Kontext spricht man in der Fachsprache auch von **Overfitting**, einer Überanpassung des Modells durch Hinzunahme irrelevanter Variablen, die zu einer unnötigen Steigerung der Komplexität führen. Das Modell scheint dabei für die vorliegenden Daten exakt zu passen, scheitert jedoch bei der Prognose von noch ungesehenen Daten. **Underfitting** ist die gegenteilige Bezeichnung und beschreibt ein zu simpel gewähltes Modell, welches zu wenig relevante Regressoren enthält und relativ schlecht an die Daten angepasst ist.[165]

2.3.2 MatLab

Da die Verfahren der Regressionsmodelle, wie im vorherigen Abschnitt beschrieben, sehr komplex und aufwendig sind, und deren Umsetzung nur noch durch rechnergestützte Mathematik möglich ist, bietet sich die Nutzung eines hierfür speziell ausgerichteten Software-Tools an.[166] Die MATLAB-Plattform stellt eine intuitive und zugleich auch interaktive Umgebung für die Modellierung von Funktionen bereit und wird zur Lösung der vorliegenden Problemstellung in der Version *R2016B*[167] verwendet. Im Folgenden soll dem Leser die Software, sowie die von ihr gebotenen Möglichkeiten zur Datenanalyse und Regressionsmodellierung, kurz vorgestellt werden.

2.3.2.1 Allgemein

MATLAB ist eine von MathWorks Inc. entwickelte Software-Plattform, die 1984 erstmalig kommerziell ausgeliefert und seitdem stetig weiter entwickelt wurde.[168,169]

[163] Vgl. *Studenmund*, Using econometrics: A practical guide, 2014, S. 56.
[164] Vgl. *Cleff*, Deskriptive Statistik und moderne Datenanalyse, 2008, S. 161.
[165] Vgl. *Cios*, Data mining: A knowledge discovery approach, 2007, S. 470.
[166] Weitere Software-Tools: R, SPSS, KNIME, SAS, uvm.
[167] Testversion für Studenten
[168] Vgl. *Pietruszka*, MATLAB in der Ingenieurpraxis, 2014, S. 1.
[169] Vgl. *Shardt*, Statistics for chemical and process engineers, 2015, S. 337.

Es wird für ML, Signal- und Bildverarbeitung, Finanzmathematik, Robotik und viele weitere Anwendungsbereiche genutzt und gilt heute als die verbreitetste Software für Analysen und Designs von Systemen und Produkten.[170] Darüber hinaus eignet es sich besonders für Data-Mining-Methoden, wie der beispielsweise Regression oder Klassifizierung. Das Portfolio umfasst rechnerunterstützte numerische Berechnungen und Visualisierungen, sowie eine eigene Hochsprache mit Programmierumgebung. Weiterhin werden vorgefertigte spezifische Bibliotheken in Form von Toolboxen angeboten, wie beispielsweise das in dieser Arbeit verwendete *Curve Fitting Tool* (*CFT*) zur Funktionsmodellierung.[171] Dadurch können beispielsweise numerisch aufwendige Verfahren der Regressionanalyse, wie die *nichtparametrische Regression* (siehe Kapitel 2.3.1.2 auf S. 31), schnell und interaktiv durchgeführt werden.

2.3.2.2 Regressionsanalyse

Die MATLAB-Plattform realisiert die Regressionsanalyse in Form der dafür vorgefertigten Bibliothek CFT. Die Bestandteile und das Verfahren dieser Toolbox soll mit Hilfe der Abbildung 14 auf S. 41 anhand eines einfachen Beispiel erläutert werden. Hierbei liegen Datenwerte zu der unabhängigen Variable x und der abhängigen Variable y vor, die zunächst den Achsen zugeordnet werden (siehe gelber Kasten oben links). Im nächsten Schritt kann aus einer Vielzahl von Funktionstypen eine passende ausgewählt werden, wie in diesem Fall eine Polynomialfunktion 1. Grades (siehe grüner Kasten), da eine *lineare Regressionsfunktion* vermutet wird. Das CFT berechnet mit den Methoden der Regressionsanalyse, sprich der Minimierung der Summe der kleinsten Quadrate (vgl. Kapitel 2.3.1 auf S. 29), die gesuchten Parameter $p1$ und $p2$ der Regressionsfunktion (siehe blauer Kasten). Des weiteren werden auch die zuvor dargestellten Bestimmtheitsmaße R^2 und $\overline{R}^2$, die den *Goodness of fit* messen, ausgeben, sodass eine Evaluation des Modells vorgenommen werden kann (siehe roter Kasten). In diesem Fall liegt mit etwa 98% eine sehr genaue Anpassung des Modells an die Daten vor. Zuletzt kann die resultierende Funktion sowie die Datenpunkte in einem Koordinatensystem betrachtet werden, um auch einen graphische Vorstellung zu erlangen. Das Tool bietet eine intuitive Benutzung, wodurch verschiedene (vermutete) Regressionsmodelle schnell und einfach mit einander verglichen werden können, ohne dabei die aufwendigen und komplexen Rechenmethoden selbst durchführen zu müssen. Neben dieser graphischen Möglichkeit bietet MATLAB auch in der mitgelieferten Programmierumgebung vorgefertigte Funktionen für die Regressionsanalyse, auf die hier jedoch nicht weiter eingegangen wird.

[170] Vgl. *Gupta*, Numerical Methods using MATLAB, 2014, S. 1.
[171] Vgl. *Pietruszka*, MATLAB in der Ingenieurpraxis, 2014, S. 1.

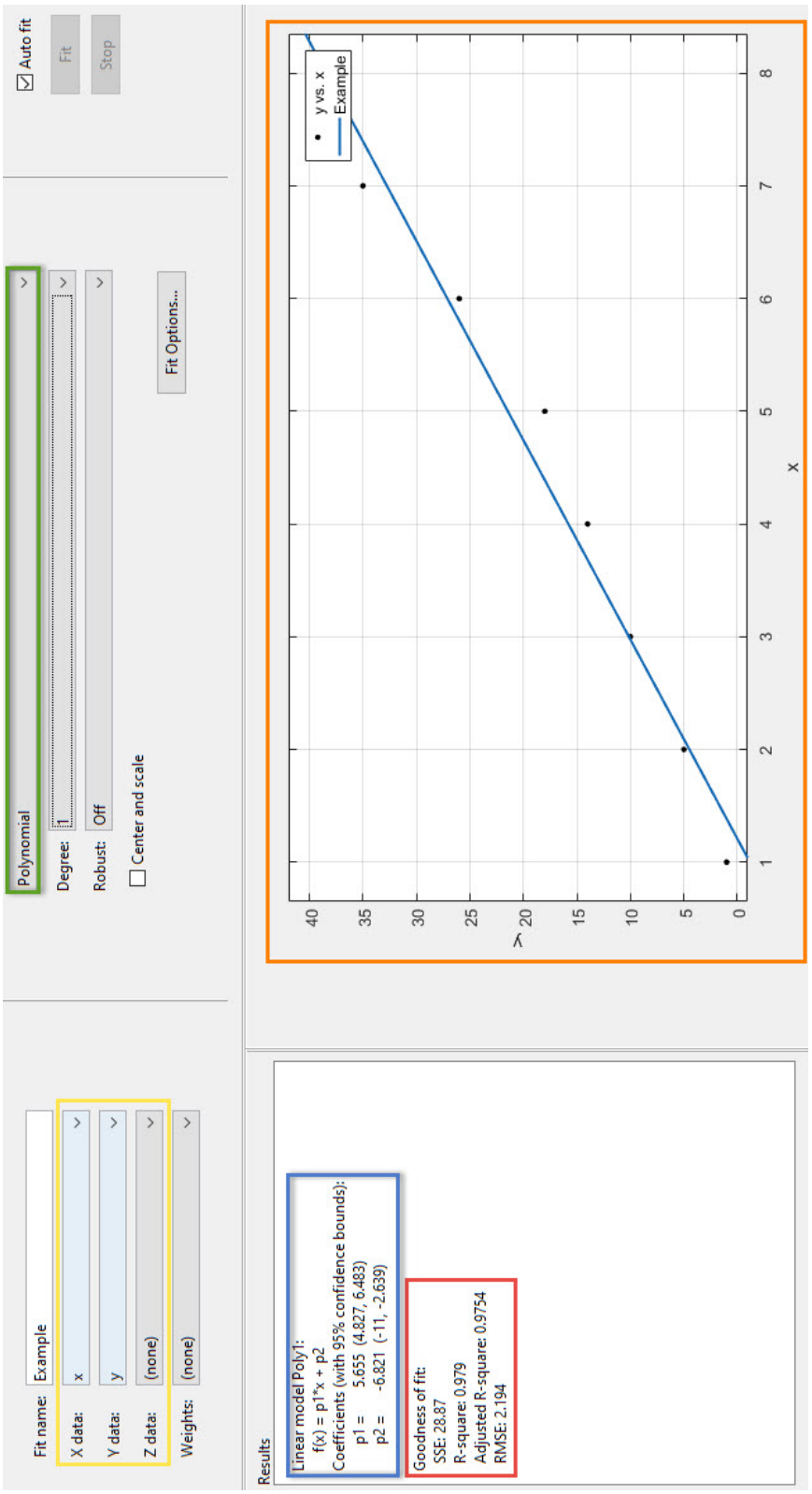

Abbildung 14: Bestandteile des Curve Fitting Tools[172]

Screenshot aus MATLAB

3 Analysephase

3.1 Expected Goals

Diese Passage soll dem Leser den aktuellen Forschungsstand der *Expected Goals* vermitteln, deren Bedeutsamkeit für den Fußballsport dabei explizit aufzeigen, sowie den Einfluss von Data-Mining-Methoden hinsichtlich der Wissensgewinnung darstellen.

„Expected Goals - Das angesagteste Statistikmodell im Profifußball“

So betitelt Nils Nordmann seinen Online-Artikel im Interview mit Dustin Böttger, Geschäftsführer von *Global Soccer Network* (*GSN*), einem der gefragtesten Datenanalysten aus Deutschland, der mit mehreren Bundesligavereinen in Kooperation steht.[173] Statistische Analysen sind im Bereich des Fußballs keine Neuheit mehr, jedoch liegt der Ursprung der sportlichen Datenanalyse in einer anderen Sportart. Der amerikanische Historiker und Statistiker Bill James veröffentlichte 1977 erste Analysen zwischen geschlagenen und gefangenen Bällen im Baseball, um eine objektive Bewertung der Gesamtleistung eines Spielers aufstellen zu können. Schumaker, Solieman und Chen bezeichnen diese Entwicklung als eine Art „Revolution“– einen Wandel von traditionellen Statistiken hin zum Wissensmanagement.[174] Diese löste eine Welle der Kreation neuer Maßzahlen aus, wovon einige im Jahr 2002 von der amerikanischen Baseball Profimannschaft *Oakland A's Billy Bean* als Grundlage zur Zusammenstellung eines neuen Teams dienten. Die *Boston Red Sox* ließen sich von dieser Idee inspirieren und gewannen anschließend sogar 2004 und 2007 die Meisterschaft.[175] Auch aus anderen Sportarten gibt es vergleichbare Beispiele, wie die Revolution im Basketball im Jahr 1980 durch den Statistiker Dean Oliver, der neue Messwerte zur Beurteilung von Spielern veröffentlichte.[176]

[173] *Nils Nordmann*, Expected Goals: Das angesagteste Statistikmodell im Profifußball, 2016.
[174] Vgl. *Schumaker/Solieman/Chen*, Sports Data Mining, 2010, S.36.
[175] Vgl. ebd.
[176] Dean Oliver beratete 2005 die *Seatlle Supersonics* und verhalf zur amerikanischen Meisterschaft

Waren im Fußball in der Vergangenheit noch rein quantitative *Key Perfomance Indicators* (*KPI*) wie der Ballbesitz, die Passquote oder die Anzahl der Torschüsse von Bedeutung, wird das Spiel heutzutage bis in das kleinste Detail (z.B. die Anzahl der vertikal „überspielten“ Gegenspieler durch einen Pass) analysiert. Durch den Fortschritt der Videotechnik können alle Aktionen eines Spieles aufgezeichnet werden, wodurch sich neue stichhaltige Bewertungsmethoden heraus kristallisiert haben. Sumpter, Anderson und weitere Fachexperten untersuchen mit Hilfe von Mathematik und Statistik das Spiel und stellen in ihren Ausführungen einige Thesen und Modelle auf.[177,178,179] Eines der momentan angesagten Modelle sind die „**Expected Goals**“(*dt. die zu erwartenden Tore*), welche die Qualität von Torschüssen vielseitig, objektiv und plausibel misst.[180] Dazu wird jedem Schuss, unter der Berücksichtigung von Parametern (wie beispielsweise der Position oder des Körperteils mit dem geschossen wurde), eine bestimmte Erfolgswahrscheinlichkeit zugewiesen. Die Bestimmung der Wahrscheinlichkeit, die Auswahl der einbezogenen Schüsse wie auch Parameter, als auch das gesamte Modell wird öffentlich von den Analytikern (meist aus Unternehmen der Sportanalyse/-beratung) kurz ausgeführt oder gar komplett geheimgehalten. Einblicke in ihre Modellierungen bieten unter anderem Opta Sports[181], der TV-Sender Sky Sports,[182] oder Experten, wie Michael Caley, in ihren Internetpublikationen.[183] Ein *Expected-Goal-Modell* offeriert eine statistisch belegte und damit objektive Bewertung von Schüssen und bildet einen neuen KPI über die Qualität einer Torchance. Anhand dieser Grundlage ist es möglich, weitere Bewertungsmethoden für Spieler und Mannschaften zu ermitteln, die vor allem im Scouting-Bereich ihre Anwendung finden. Durch die qualitative Bewertung der Schüsse eines Stürmers mittels des Expected-Goal-Modells, kann eine objektive Aussage über dessen Erfolgsquote getroffen werden (beispielsweise ob diese über den erwarteten Tore liegt), welche dann zur Spielersuche herangezogen werden kann. Eine Gefahr in der Modellierung der Expected Goals stellt die *Überparametrisierung* (vgl. Kapitel 2.3.1.3 auf S. 38) dar. Werden zu viele Parameter, z.B. welcher Spieler geschossen hat und ob mit seinem starken oder schwachen Fuß geschossen wurde, seine Tagesform, die Leistung des generischen Torhüters, usw. für die Modellierung herangezogen, verliert das Modell durch zu viele Details seine Abstraktion und folglich die allgemeine Aussagekraft für alle Schüsse. Die Kunst liegt in der Kapitel 2.3.1.3 auf S. 38 beschriebenen Balance von *Underfitting* und *Overfitting* des Modells.

[177] Vgl. *Sumpter*, Soccermatics, 2016.
[178] Vgl. *Anderson/Sally*, The numbers game, 2014.
[179] Vgl. *Heuer/Müller/Rubner*, Soccer: Is scoring goals a predictable Poissonian process?, 2010.
[180] *Nils Nordmann*, Expected Goals: Das angesagteste Statistikmodell im Profifußball, 2016.
[181] Vgl. *Philipp Obloch*, OptaPro Neuheit: Expected Goals, 2015.
[182] Vgl. *Philipp Ertl*, Expected Goals: Welche Teams sind am effizientesten?, 2016.
[183] Vgl. *Michael Caley*, Bringing baseball stat nerdiness to football, 2017.

Durch die Professionalisierung der Datenaufnahme im Fußball werden stetig mehr Daten während eines Spieles erhoben[184], woraus im Laufe einer Saison eine Datenmenge resultiert, die die Leistungsfähigkeit herkömmlicher Analysewerkzeuge übersteigt. Um wertvolle Informationen aus den umfangreich Daten zu extrahieren, greifen auch Datenanalysten im Bereich des Fußball auf die Prozesse und Methoden des Data Minings zurück. Ausführliche Einblicke in die Komplexität der Datenanalyse im Sport stellen unter anderem Schumaker et al. in ihrer Ausarbeitung vor.[185] Data-Mining-Methode wie das Clustering zur Einteilung von Spielertypen, die Regressionanalyse zur Ermittlung von Erfolgsfaktoren einer Saison, Entscheidungsbäume zur Bestimmung des perfekten Ein- und Auswechslungszeitpunktes, als auch Neuronale Netze zur Prognose von Spielausgängen, werden zur Wissensgewinnung verwendet.[186] Darüber hinaus werden einige dieser Techniken zur komplexen Erkennung von Taktiken und Spielphilosophien eingesetzt, welche in der Ausführung von Rein konkretisiert werden.[187]

3.2 Opta-Spieldaten

Im Folgenden Abschnitt wird der Leser mit den zugrundeliegenden Daten dieser Arbeit vertraut gemacht, welche die Basis für die anschließende Funktionsmodellierung bilden. Es wird ein Überblick über das Format der bereitgestellten Daten, als auch der darin enthaltenden Informationen gegeben, um die in der Umsetzung verwirklichten Prozessschritte der Datenselektion, -vorverarbeitung und -transformation nachvollziehen zu können.

Neben Datenprovidern wie *Die Liga – Fußballverband e.V.* (*DFL*), Heimspiel oder Amisco, liefert der weltweit führende Anbieter von Sportdaten, *Opta-Sports*, unter Anderem detaillierte Informationen über Spielevents. Dabei werden pro Spiel zwischen 1600 bis 2000 Aktionen, darunter Pässe, Fouls, Tore uvm., erfasst und aufbereitet.[188] Opta stellt die Daten mittels einer *Extensible Markup Language* (*XML*)-Datei (siehe Abbildung 21 auf S. 59) bereit, welche zunächst für eine bessere Weiterverarbeitung in ein *JavaScript Object Notation* (*JSON*)-Format *geparst* werden. Listing 1 auf S. 45 zeigt dazu exemplarisch ein Opta-Event mit allen darin beinhaltenden Informationen.

[184] beispielsweise durch Videobildverarbeitung oder Sensordaten
[185] Vgl. *Schumaker/Solieman/Chen*, Sports Data Mining, 2010.
[186] Vgl. *Gunjan Kumar*, Machine Learning for Soccer Analytics,.
[187] Vgl. *Rein/Memmert*, Big data and tactical analysis in elite soccer, 2016.
[188] Vgl. *Opta-Sports*, The collection process, 2017a.

```
1  {
2    "$": {
3      "id": "2016516030",
4      "event_id": "1029",
5      "type_id": "16",
6      "period_id": "1",
7      "min": "36",
8      "sec": "31",
9      "player_id": "55634",
10     "team_id": "156",
11     "outcome": "1",
12     "x": "95.0",
13     "y": "43.5",
14     "timestamp": "2014-08-22 20:14:39.71",
15     "last_modified": "2014-08-25 14:25:09"
16   },
17   "Q": {
18       ...
19   }
20 }
```

Listing 1: Struktur der Opta-Daten

Jedes Event besitzt eine eindeutige ID innerhalb des Spiels („event_id“) und eine spezifische, dem Spiel übergeordnete „id“, welche zur Identifizierung in der gesamten Opta-Datenbank dient. Über die „player_id“ kann der agierende Spieler des Events identifiziert werden, sowie dessen Teamzugehörigkeit über die „team_id“. Des Weiteren kann die Position der Aktion auf dem Spielfeld über die „x“- und „y“-Koordinaten lokalisiert, als auch der genau Zeitpunkt ermittelt werden (siehe Zeile 6-8). Das Attribut „type_id“ beschreibt die Art eines Events durch einen numerischen Wert, wobei in diesem Fall die 16 für einen Torerfolg steht. Eine „type_id“ von 1 beispielsweise würde einen Pass repräsentieren, während der „outcome“ den Erfolg des Events angibt. Ein *„outcome“* von 1 wäre dann ein erfolgreicher Pass zum Mitspieler, ein „outcome“ von 0 hingegen würde einen Fehlpass widerspiegeln. Unter dem Attribut „Q“ finden sich über 300 mögliche Qualifier, welche weitere detaillierte Informationen zu bestimmten Events geben. Die Qualifier 9 bzw. 28 geben beispielsweise Auskunft, ob der Torerfolg aus einem Elfmeter oder aus einem Eigentor resultierte.[189]

Die Position des Events wird wie beschrieben durch die „x“- und „y“-Koordinaten bestimmt und kann mit Hilfe der Abbildung 15 auf S. 46 graphisch aufgezeigt werden. Die eigene Torlinie liegt dabei immer bei x=0, die gegnerische Torlinie bei

[189] Vgl. *Opta-Sports*, F24 Appendices, 2017b, S.1.

x=100, sowie die Mitte des gegnerischen Tores bei y=50, wodurch sich die Spielrichtung von links nach rechts ergibt. Der in Listing 1 auf S. 45 dargestellte Schuss wurde folglich von einer zum Tor relativ nahen Position ausgeführt. Die Werte für x und y werden dabei prozentual zur Spielfeldlänge bzw. -breite angeben, um Events aus verschiedenen Spielen (=unterschiedliche Spielfeldgrößen) vergleichen zu können.[190] Der spätere Ursprungspunkt der Funktion (x=0 und y=0) soll in der Mitte des gegnerischen Tores (x=100 und y=50) liegen, sodass die Koordinatenpunkte transformiert werden müssen (siehe Anforderung 6).

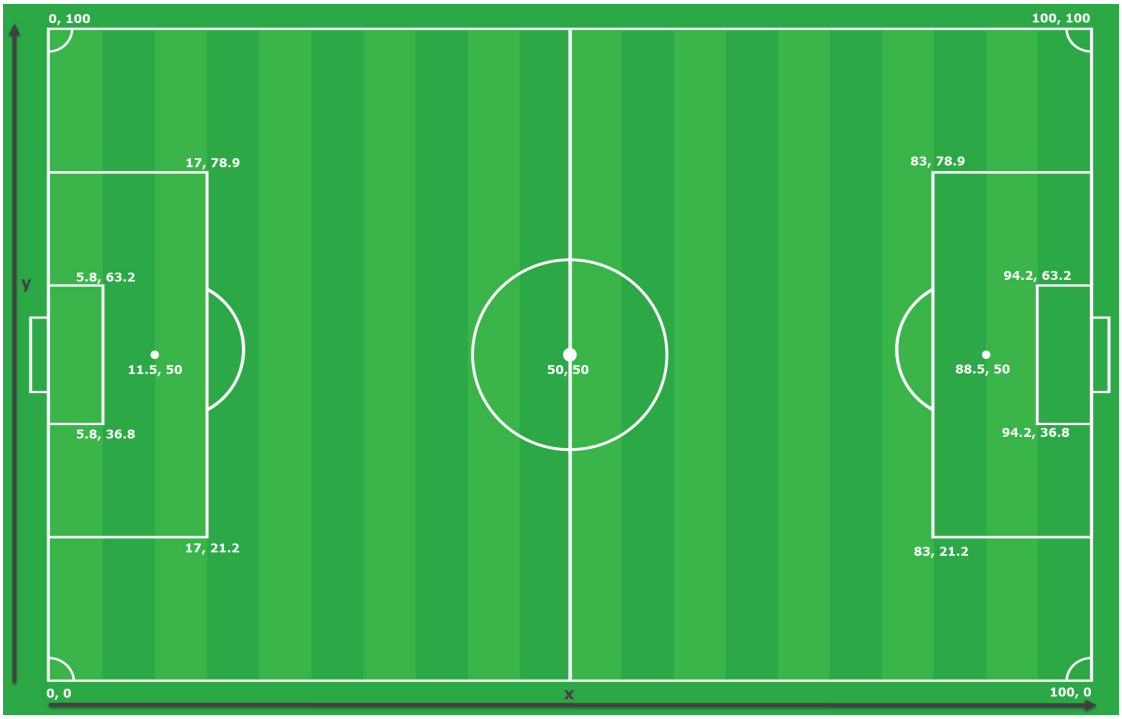

Abbildung 15: Koordinatensystem Opta[191]

Die vorliegende Arbeit beschäftigt sich mit der Wahrscheinlichkeit eines Torerfolges von einer bestimmten Position. Folglich liegt der Fokus auf der Selektion aller Schüsse, die über die Type-IDs 13, 14, 15 und 16 identifiziert werden können. Tabelle 2 auf S. 47 gibt dazu einen Überblick über die einzelnen Events inklusive kurzer Beschreibungen. Aus der Tabelle können die Schüsse in die für die Funktionsmodellierung wichtigen Kategorien **Tor** und **Nicht-Tor** eingeteilt werden. Schüsse mit der Type-ID 13, 14, 15 werden demzufolge als Nicht-Tor sowie die Type-ID 16 als Tor verbucht.

[190] Vgl. *Opta-Sports*, F24 Appendices, 2017b, S.1.
[191] Abbildung in Anlehnung an *Opta-Sports*, F24 Appendices, 2017b, S. 43.

Schuss-Events		
Type-ID	**Name**	**Beschreibung**
13	Miss	Jeder Schuss, der am Tor vorbei ging
14	Post	Der Ball hat den Torrahmen getroffen
15	Attempt Saved	Alle Schüsse, die gehalten wurden
16	Goal	Alle Tore

Tabelle 2: Schuss-Events

Aus der dritten Anforderung (vgl. Auflistung der Anforderung Tabelle 1 auf S. 4) lässt sich schließen, dass für die Modellierung nur Schüsse berücksichtigt werden dürfen, die aus dem „*laufenden*“ Spiel abgegeben wurden. Die Qualifier 9, 25 und 26 geben darüber Aufschluss, ob ein Schuss aus einer Standardsitutation (Eckball, Freistoß oder Elfmeter) hervorgeht. Ferner müssen Eigentore ausgeschlossen werden, da diese – wie in Anforderung 4 beschrieben – die Modellierung verzerren würden. Geblockte Schüsse müssen ebenfalls unberücksichtigt bleiben, anlässlich der Tatsache, dass keine konkrete Aussage getroffen werden kann, ob solch ein Schuss in einem Torerfolg resultiert hätte (vgl. Anforderung 5). Tabelle 3 fasst dazu nochmal alle relevanten Qualifier zusammen.

Relevante Qualifier		
Q-ID	**Name**	**Beschreibung**
9	Penalty	Schussversuch resultiert aus einem Elfmeter
25	From Corner	Schussversuch resultiert direkt aus einer Ecke
26	Free Kick	Schussversuch resultiert direkt aus einem Freistoß
28	Own Goal	Eigentor (inverse Koordinaten)
82	Blocked	Schussversuch der geblockt wurde

Tabelle 3: Relevante Qualifier

Als Datenbasis für diese Arbeit wurden von Opta folgende Spieldaten aus der ersten deutschen Fußballliga bereitgestellt:

- Bundesliga-Saison 2013/2014 (GER)
- Bundesliga-Saison 2014/2015 (GER)
- Bundesliga-Saison 2015/2016 (GER)
- Bundesliga-Saison 2016/2017 (GER)

Damit liegt eine ausreichend große und repräsentative Datenmenge mit über 15.000 verfügbaren Schüssen für eine Funktionsmodellierung vor, die im folgenden Kapitel (vgl. Kapitel 4 auf S. 49) umgesetzt wird.

3.3 Wirtschaftliche Betrachtung

- Scouting
- Einkauf von Daten
- Zusammenstellung einer ganzen Mannschaft (siehe Baseball, Basketball)
- ...

4 Umsetzung

Die Umsetzung der vorliegenden Problemstellung wird mit Hilfe des in Kapitel 2.2 auf S. 13 vorgestellten KDD-Prozesses durchgeführt. Anhand der erlangten Grundlagen und Methoden der einzelnen Prozessschritte, werden diese sukzessive durchlaufen, um eine möglichst exakte Modellierung der Funktion zu realisieren. Zunächst werden die relevanten Daten in Kapitel 4.1 selektiert, anschließend aufbereitet (vgl. Kapitel 4.2 auf S. 51) und in das für die Regressionsanalyse passende Format transformiert (vgl. Kapitel 4.3 auf S. 54). Im darauf folgenden Schritt des Data Minings wird die Funktion unter der Betrachtung des Winkels und der Distanz, als auch in Bezug auf die Koordinaten des Schusses anhand der Regression (vgl. Kapitel 2.3 auf S. 29) modelliert. Abschließend werden die aus MATLAB gewonnen Ergebnisse interpretiert und evaluiert, um eine fundierte Entscheidung über die Auswahl eines Modells treffen zu können.

4.1 Datenselektion

Opta Sports erfasst in einem Fußballspiel zwischen 1600 und 2000 verschiedene Events (z.B. Pässe, Schüsse, Fouls, uvm.), wovon eine geeignete Datenmenge für die Funktionsmodellierung ausgewählt werden muss. Zunächst werden die im XML-Format vorliegenden Daten eines Spieles für eine bessere Weiterverarbeitung in ein JSON-Format geparst. Über die in Kapitel 3.2 auf S. 44 beschriebenen Type-IDs und Qualifiers können die relevanten Schüsse in dieser großen Datenmenge selektiert werden. Aus der Tabelle 2 auf S. 47 lassen sich alle Schussversuche innerhalb eines Spiels identifizieren, wobei die Zieldaten nur Schüsse beinhalten dürfen, die während des „freien“ Spiels abgeben wurden, nicht aus Eigentoren resultierten und welche die nicht geblockt wurden (vgl. Tabelle 1 auf S. 4). Der Ausschluss solcher Schüsse erfolgt über die in Tabelle 3 auf S. 47 gelisteten Qualifier. Die Eigentore wurden nicht von vornherein ausgeschlossen und konnte erst durch die Visualisierung aller Torerfolge in Abbildung 16 auf S. 50 als irrelevant erkannt werden, da diese die Wahrscheinlichkeit ein Tor aus dieser Position zu erzielen total verzerren. Das „Eigentor des Jahres“[192] von Christopher Kramer aus 45-Meter kann beispielsweise ohne Angabe des Qualifiers auch als sehr weiten Fernschuss interpretiert werden.

[192] Vgl. *F.A.Z.*, Das Eigentor des Jahres, 2014.

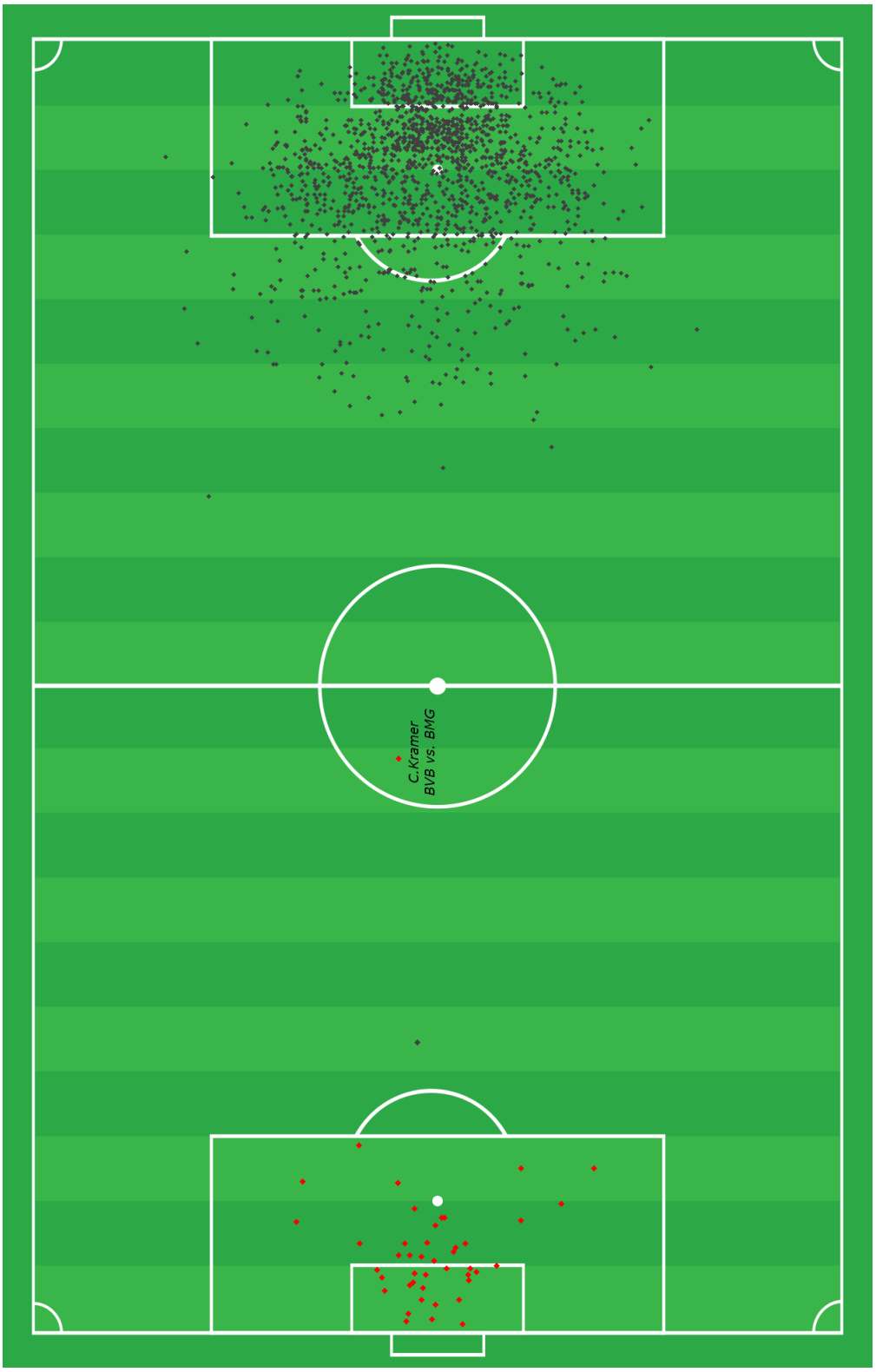

Abbildung 16: Ausschluss der Eigentore (rote Punkte)

4.2 Datenvorverarbeitung

Wie in den Grundlagen der Datenvorverarbeitung (vgl. Kapitel 2.2.2 auf S. 14) begründet, ist diese Phase von besonderer Bedeutung für die Güte der DM-Resultat. Es gilt die Qualität der zuvor lediglich selektierten Zieldaten durch den Einsatz von geeigneten Verfahren nachhaltig zu verbessern und diese in einem für die DM-Methode passendes Format bereitzustellen. Dazu müssen im Folgenden die Daten (in Form der selektierten Schüsse) zunächst von Fehlern bereinigt (siehe Data Cleaning **??** auf S. ??), anschließend fehlerfrei mit allen anderen verfügbaren Spiele zusammengeführt (siehe Data Integration Kapitel 4.2.2 auf S. 52)) und schließlich auf eine verwertbare Datenmenge reduziert werden (siehe Data Reduction Kapitel 4.2.3 auf S. 52).

4.2.1 Data Cleaning

Das Data Cleaning beschäftigt sich mit den Problemarten der fehlenden, verrauschten und inkonsistenten Daten (vgl. Kapitel 2.2.2.1 auf S. 16), welche nachfolgend näher untersucht und behandelt werden.

Fehlende Daten Das ein ganzes Event, wie in diesem Falle ein Schuss, überhaupt nicht erfasst wurde, ist sehr unwahrscheinlich, da Opta als markführender Datenprovider großen Wert auf Akkuratesse und Vollständigkeit legt. Für eine exakte Überprüfung müssten über 90.000 Spielminuten (*34 Spieltage x 3,5 Saisons x 9 Spiele pro Spieltag x 90 Minuten* darauf händisch geprüft werden, was zu einem nicht realisierbaren Vorgang führt. Fehlen „lediglich“einzelne Attribute in einem Datensatz, wie beispielsweise die x-Koordinate des Schusses, können diese Datensätze auf unterschiedliche Weise behandelt werden (siehe Kapitel 2.2.2.1 auf S. 16).

verrauschte Daten und Ausreißer Outlier-Detection → Stoppelkamp Tor[193]

inkonsistente Daten Behandlung fehlerhafter Daten (z.B. $x > 100$)

Das weiße Linienproblem

[193] Vgl. *DFB*, Rekordtore im deutschen Fußball, 2014.

4.2.2 Data Integration

- Konkatenieren aller Spiele
- Einheitliches Format

4.2.3 Data Reduction

- Irrelevante Attribute entfernen
- Unterscheidung zwischen Tor(1) und Nicht-Tor(0)
- Ziel: {"x":_, "y":_, "goal":_}

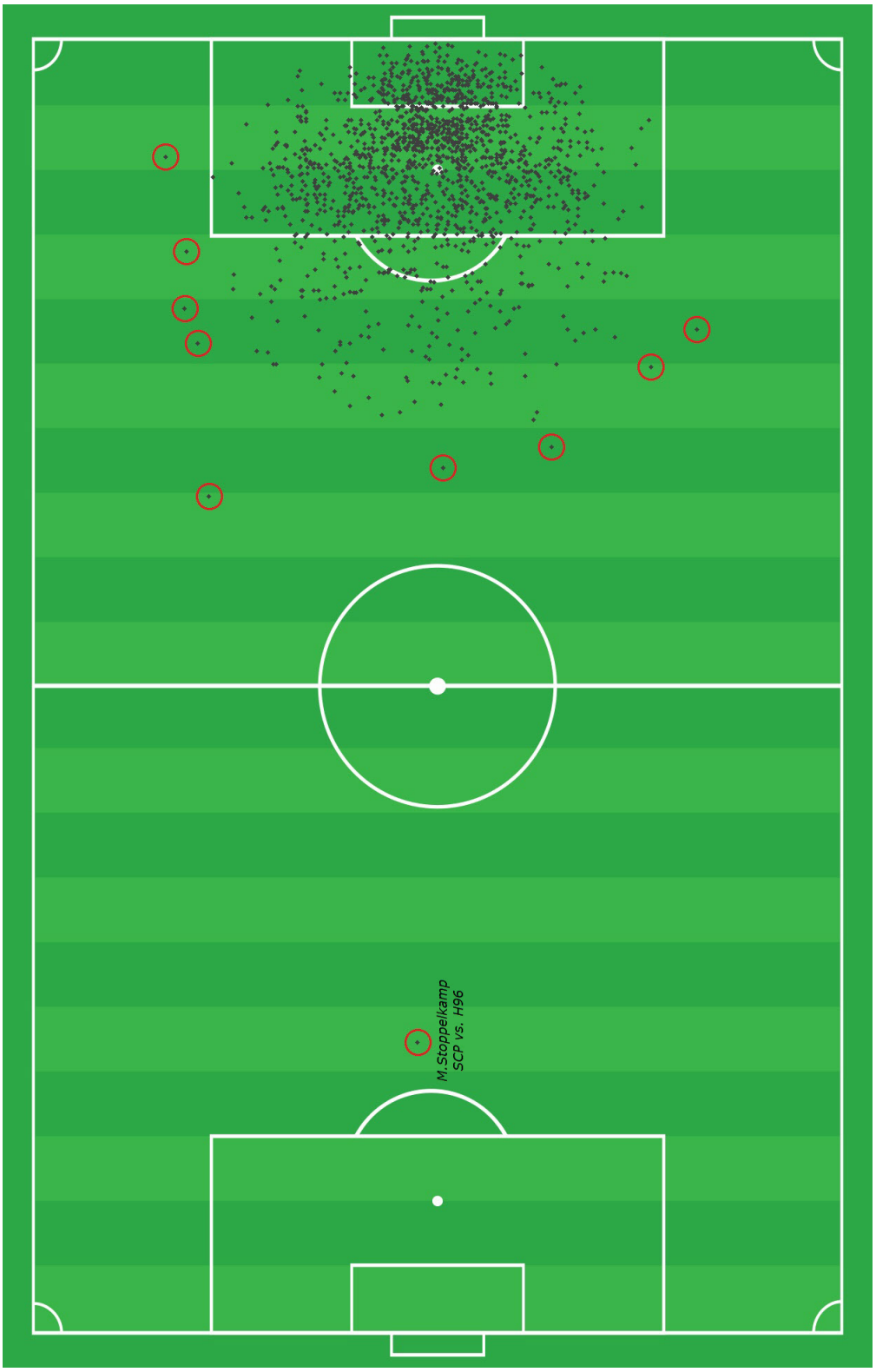

Abbildung 17: Darstellung der Ausreißer bei Torerfolgen

4.3 Datentransformation

4.3.1 Allgemeine Transformationen

- Transformieren des Koordinatensystems → Abbildung
- Seiten des Spielfeldes an y-Achse ($x = 0$) spiegeln
- Zielformat für MATLAB vorbereiten $(x|y|p)$

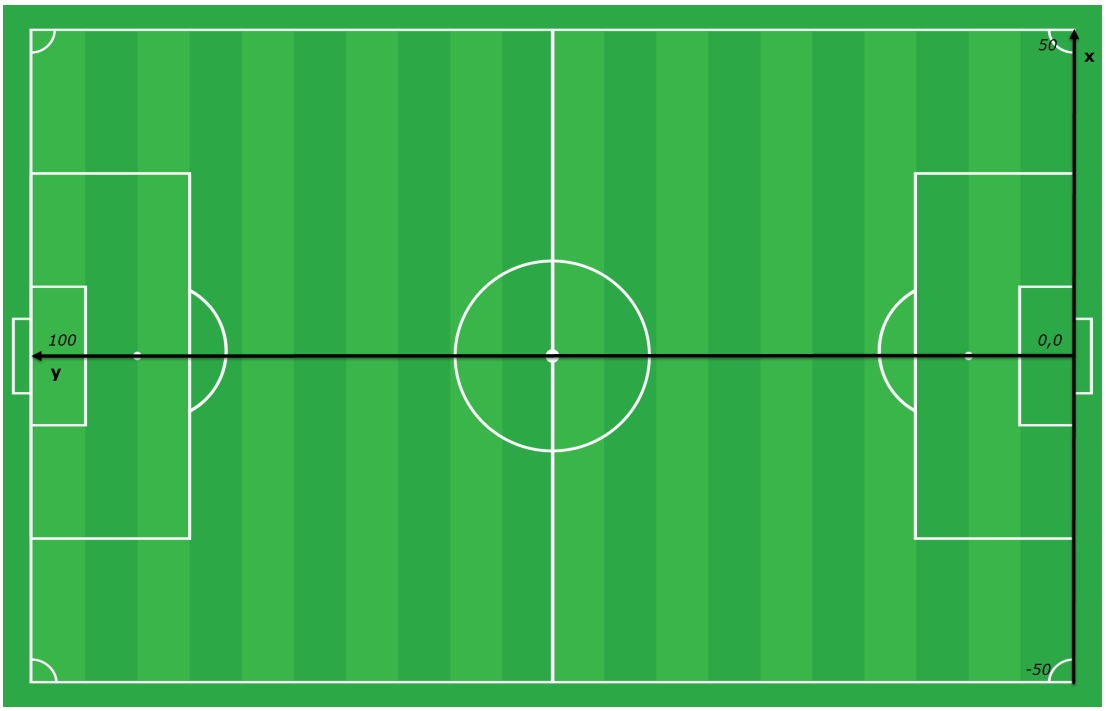

Abbildung 18: Transformation des Koordinatensystems

4.3.2 Transformation für Winkel- und Distanzbetrachtung

- Winkel berechnen
- Distanz berechnen
- Einteilung in Intervalle

- Abbildung der Berechnung

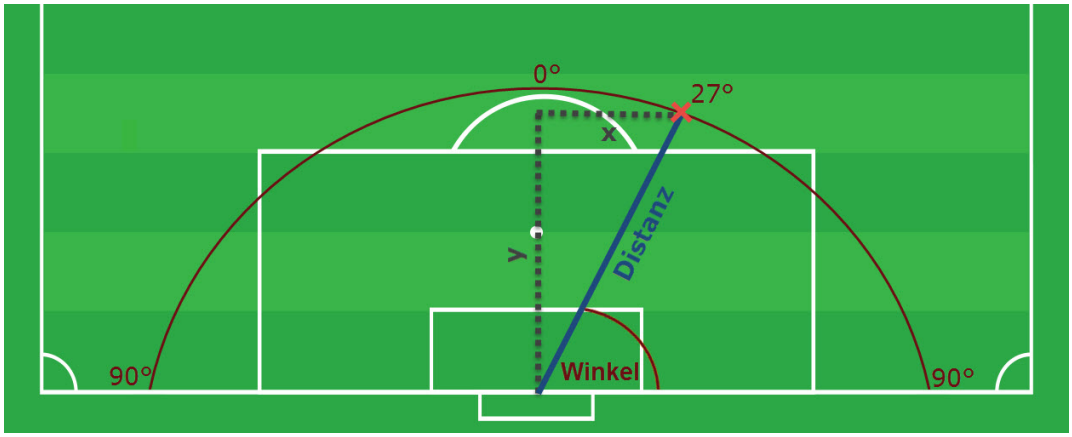

Abbildung 19: Berechnung der Distanz und des Winkels

$$Distanz = \sqrt{x^2 + y^2} \tag{11}$$

$$Winkel = \arctan(\frac{|x|}{y}) \tag{12}$$

4.3.3 Transformation für Koordinatenbetrachtung

- Spielfeld in Quadrate einteilen → Abbildung
- Wahrscheinlichkeit (p) für jedes Quadrat berechnen → $\frac{Anzahl\ der\ Tore\ im\ Quadrat}{Gesamtzahl\ der\ Schüsse\ im\ Quadrat}$
- Seiten des Spielfeldes an y-Achse ($x = 0$) spiegeln → Durchschnitt der Wahrscheinlichkeiten
- Zielformat für MATLAB vorbereiten ($x|y|p$)

$$p = \frac{Anzahl\ der\ Tore\ im\ Quadrat}{Gesamtzahl\ der\ Schüsse\ im\ Quadrat} \tag{13}$$

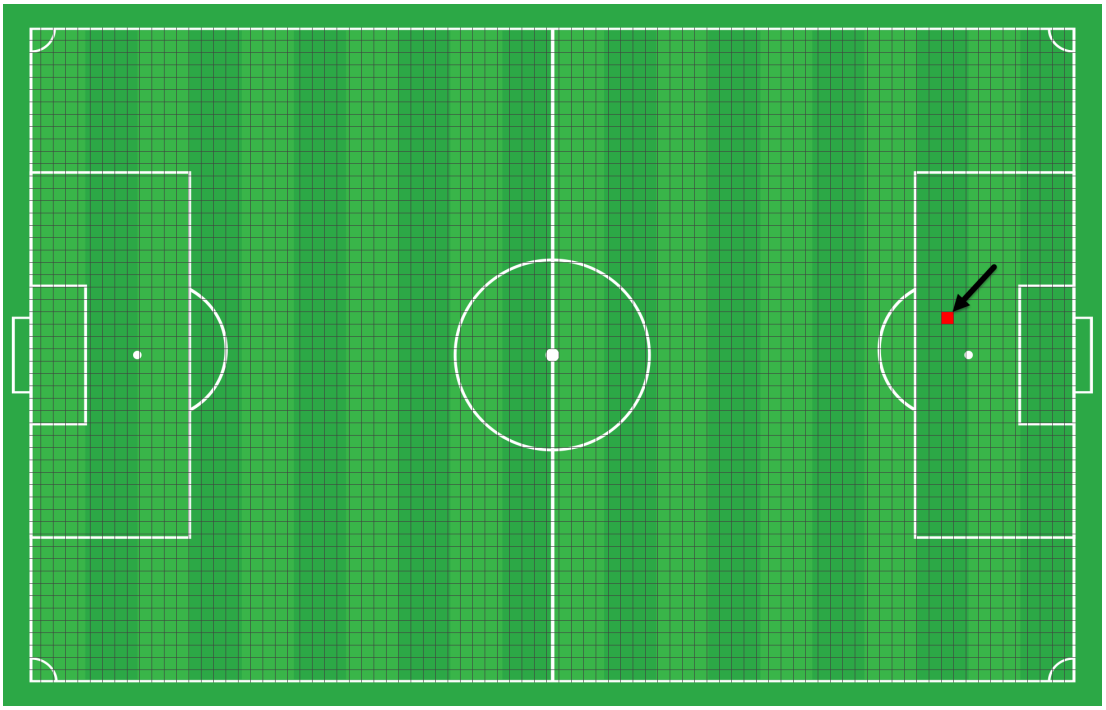

Abbildung 20: Einteilung des Spielfeldes in Raster

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 Opta

Abbildung 21: Auszug aus der XML-Datei mit den Events

Die Abbildung 21 zeigt einen Ausschnitt aus der XML-Datei in der alle Events eines Spiels aufzeichnet sind. Im oberen Teil sind die Mannschaftsaufstellungen zu sehen, wobei jeder Spieler eine eigene ID besitzt. Darunter folgt gelb markiert zum Anpfiff der Partie der erste Pass mit dem *Outcome* 1, welcher einen erfolgreichen Pass identifiziert.

Abbildung 22: Problem der weißen Linien

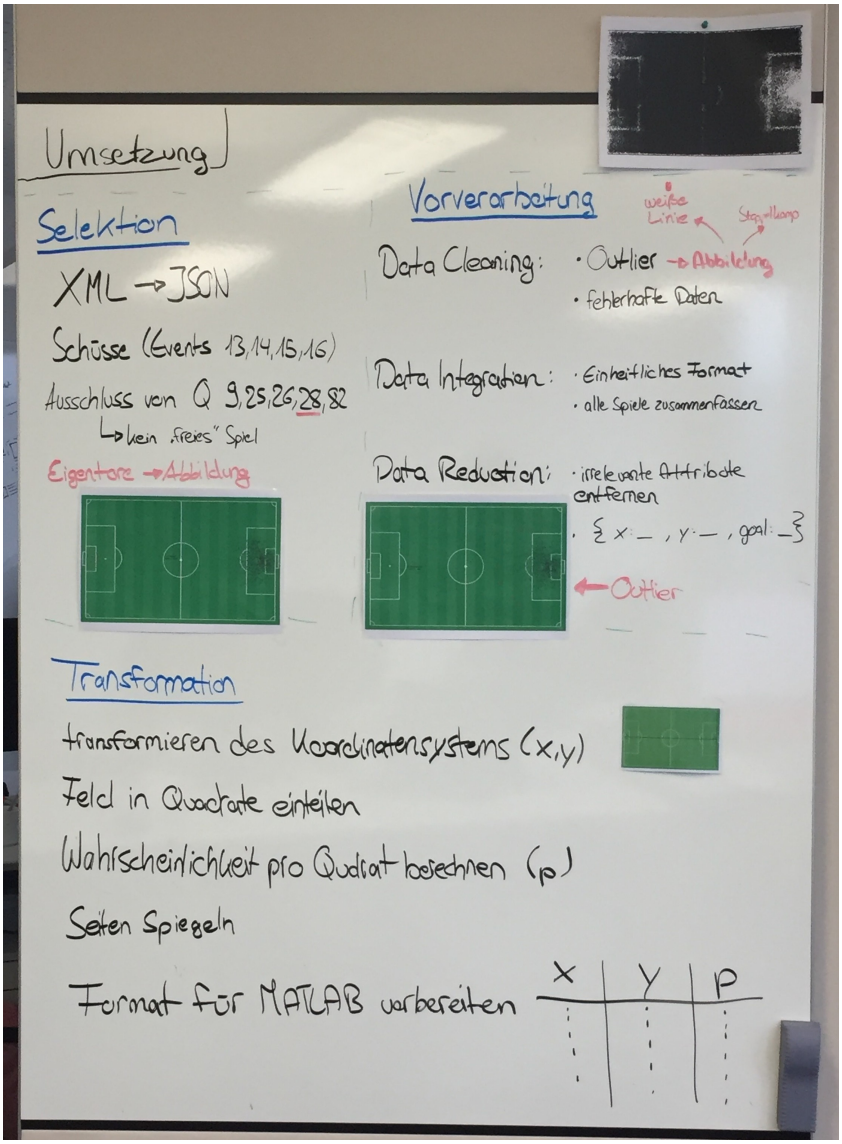

Abbildung 23: Überblick Umsetzung (wird wieder rausgenommen)

B MatLab Code

Glossar

Big Data
Definition folgt → S. 7

Clustering
Definition folgt → S. 18

Cross Industry Standard Process for Data Mining (CRISP-DM)
Definition folgt → S. 9, 10, 12

Data Mining (DM)
Definition folgt → S. 6, 9, 14, 22, 24, 51

Internet of Things (IoT)
Definition folgt → S. 7

K-Nearest Neighbours (KNN)
Definition folgt → S. 17, 22

Knowledge Discovery in Data (KDD)
Definition folgt → S. 9, 12, 13, 24, 28

Künstliche Intelligenz
Definition folgt → S. 5, 25

Machine Learning (ML)
Definition folgt → S. 5, 25, 40

Methode der kleinsten Quadrate (MDKQ)
Definition folgt → S. 30

Neuronale Netze (NN)
Definition folgt → S. 25

Online Analytical Processing (OLAP)
Definition folgt → S. 13

Regression
Definition folgt → S. 18, 29

Structured Query Language (SQL)
Definition folgt → S. 13

Super Vector Machine (SVM)
Definition folgt → S. 25

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

Anderson, Chris/Sally, David [The numbers game, 2014]: The numbers game: Why everything you know about football is wrong. [fully updated with new world cup chapter] Auflage. London u.a.: Penguin Books, 2014, ISBN 978–0–241–96362–3

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

Cios, Krzysztof J. [Data mining: A knowledge discovery approach, 2007]: Data mining: A knowledge discovery approach. New York and New York: Springer, 2007, ISBN 0387333339

Cleff, Thomas [Deskriptive Statistik und moderne Datenanalyse, 2008]: Deskriptive Statistik und moderne Datenanalyse: Eine computergestützte Einführung mit Excel, SPSS und STATA ; [Bachelor geeignet! Lehrbuch]. 1. Auflage. Wiesbaden: Gabler, 2008, ISBN 978–3–8349–0202–3

Cleve, Jürgen/Lämmel, Uwe [Data Mining, 2014]: Data Mining. [Elektroni-

sche Ressource] Auflage. München: De Gruyter Oldenbourg, 2014, ISBN 9783486713916

Daróczi, Gergely [Mastering data analysis, 2015]: Mastering data analysis with R: Gain clear insights into your data and solve real-world data science problems with R - from data munging to modeling and visualization. Birmingham: Packt Publishing, 2015, Community experience distilled, ISBN 9781783982028

DFB [Rekordtore im deutschen Fußball, 2014]: Stoppelkamp und wer noch? Rekordtore im deutschen Fußball. 2014 ⟨URL: `http://www.dfb.de/news/detail/stoppelkamp-und-wer-noch-rekordtore-im-deutschen-profifussball-106319/`⟩ – Zugriff am 30.01.2017

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

F.A.Z. [, 2014]: Das Eigentor des Jahres. 2014 ⟨URL: `http://www.faz.net/aktuell/sport/fussball/bundesliga/kramer-schiesst-das-eigentor-des-jahres-13258164.html`⟩ – Zugriff am 30.01.2017

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

Gunjan Kumar [Machine Learning for Soccer Analytics,]: Machine Learning for Soccer Analytics. ⟨URL: `https://www.researchgate.net/publication/257048220?channel=doi&linkId=0c96052441dfabfc87000000&showFulltext=true`⟩ – Zugriff am 09.01.2017

Günther, Marco/Velten, Kai [Mathematische Modellbildung und Simulation, 2014]: Mathematische Modellbildung und Simulation: Eine Einführung für Wissenschaftler, Ingenieure und Ökonomen. Weinheim: Wiley-VCH-Verl., 2014, Lehrbuch Physik, ISBN 978–3–527–41217–4

Gupta, Abhishek K. [Numerical Methods using MATLAB, 2014]: Numerical Methods using MATLAB. Berkeley CA: Apress, 2014, ISBN 978–1–4842–0154–1

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

Hastie, Trevor J./Tibshirani, Robert J./Friedman, Jerome H. [The elements of statistical learning, 2016]: The elements of statistical learning: Data mining, inference, and prediction. 2. Auflage. New York, NY: Springer, 2016, Springer series in statistics, ISBN 9780387848570

Heuer, A./Müller, C./Rubner, O. [Soccer: Is scoring goals a predictable Poissonian process?, 2010]: Soccer: Is scoring goals a predictable Poissonian process? EPL (Europhysics Letters), 89 2010, Nr. 3, 38007 ⟨URL: `https://www.researchgate.net/publication/45899136_Soccer_Is_scoring_goals_a_predictable_Poissonian_process`⟩, ISSN 0295–5075

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

Michael Caley [, 2017]: Bringing baseball stat nerdiness to football. 2017 ⟨URL: `https://mcofa.wordpress.com/`⟩ – Zugriff am 27.01.2017

Nils Nordmann [Expected Goals: Das angesagteste Statistikmodell im Profifußball, 2016]: Expected Goals: Das angesagteste Statistikmodell im Profifußball. 2016 ⟨URL: `https://www.welt.de/sport/fussball/article151870094/Das-angesagteste-Statistikmodell-im-Profifussball.html`⟩ – Zugriff am 20.01.2017

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

Philipp Ertl [Expected Goals: Welche Teams sind am effizientesten?, 2016]: Expected Goals: Welche Teams sind am effizientesten? 2016 ⟨URL: `http://www.skysportaustria.at/bundesliga-at/expected-goals-welche-teams-sind-am-effizientesten/`⟩ – Zugriff am 26.01.2017

Philipp Obloch [OptaPro Neuheit: Expected Goals, 2015]: OptaPro – Neuheit: Expected Goals. 2015 ⟨URL: `http://www.optasportspro.com/de/ueber-uns/optapro-blog/posts/2015/blog-optapro-%E2%80%93-neuheit-expected-goals/`⟩ – Zugriff am 26.01.2017

Pietruszka, Wolf Dieter [MATLAB in der Ingenieurpraxis, 2014]: MATLAB® und Simulink® in der Ingenieurpraxis: Modellbildung, Berechnung und Simulation. 4. Auflage. Springer Fachmedien Wiesbaden, 2014, ISBN 978–3–658–06420–4

Rein, Robert/Memmert, Daniel [Big data and tactical analysis in elite soccer, 2016]: Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus, 5 2016, Nr. 1, 1410

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

Schimek, Michael G. (Hrsg.) [Smoothing and regression, 2000]: Smoothing and regression: Approaches, computation, and application. New York NY u.a.: Wiley, 2000, Wiley series in probability and statistics : Applied probability and statistics section, ISBN 0–471–17946–9

Schumaker, Robert P./Solieman, Osama K./Chen, Hsinchun [Sports Data Mining, 2010]: Sports Data Mining. Band 26, Integrated Series in Information Systems. 1. Auflage. Boston MA: Springer Science+Business Media LLC, 2010, ISBN 978–1–4419–6730–5

Shardt, Yuri A.W. [Statistics for chemical and process engineers, 2015]: Statistics for chemical and process engineers: A modern approach. 2015, ISBN 978–3–319–21509–9

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

Studenmund, Arnold H. [Using econometrics: A practical guide, 2014]: Using econometrics: A practical guide. Pearson new international ed. of 6th rev. ed. Auflage. Boston Mass. u.a.: Pearson, 2014, ISBN 978–1–29202–127–0

Sumpter, David [Soccermatics, 2016]: Soccermatics: Mathematical adventures in the beautiful game. London: Bloomsbury Sigma, 2016, ISBN 1472924134

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