UNIVERSITÄT HOHENHEIM

Freie wissenschaftliche Arbeit zur Erlangung des akademischen Grades
Master of Science

Anwendung von Data-Mining-Technologien zu statistischen Auswertungen und Vorhersagen im Fußball

Prof. Dr. Mareike Schoop

Fakultät Wirtschafts- und Sozialwissenschaften (500)
Institut für Interorganisational Management & Performance (580)
Lehrstuhl für Wirtschaftsinformatik 1 (580A)
Universität Hohenheim

Betreuer: Michael Körner

Andreas Brauchle
Röschbühlweg 7
88339 Bad Waldsee

Matrikelnummer: 601700

Tel.: 0176/60996234
Email: andreas.brauchle@uni-hohenheim.de

Master Wirtschaftsinformatik

Stuttgart, 05. Oktober 2016

Wir sind Lichtjahre davon entfernt zu wissen, wie der Fußball funktioniert, und es trennen und sogar ganze Galaxien davon zu wissen, wie Erfolg im Fußball zustande kommt

Dr. Roland Loy, Sportwissenschaftler und Fußballstatistiker (Loy, 2008, S. 15)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BI ..... Business Intelligence

CRISP-DM ..... Cross Industry Standard Process for Data Mining

DFB ..... Deutscher Fußball Bund e.V.

DFL ..... Deutsche Fußball Liga GmbH

DWH ..... Data Warehouse

KDD ..... Knowledge Discovery in Databases

KPI ..... Key Performance Indicator

SVM ..... Support Vector Machine

1 Einleitung

In Unternehmen werden heutzutage sehr große Datenmengen erfasst und verarbeitet, wobei diese Daten aus den verschiedensten Quellen stammen können. Neben internen Daten (z. B. Kundendaten, Bestellungen, uvm.) können die Unternehmen auch auf externe Datenquellen zurückgreifen (z. B. Daten aus dem Internet, Daten von verbundenen Unternehmen etc.). Auf der Grundlage solcher heterogenen Datenmengen wird versucht, durch Datenanalysen – beispielsweise durch Data Mining – aussagekräftiges Wissen zu generieren, das im Rahmen des operativen Geschäftes für Managementaufgaben und zur Entscheidungsunterstützung verwendet werden kann, um im optimalen Fall einen Wettbewerbsvorteil gegenüber Konkurrenten zu generieren.

Die Deutsche Bundesliga besteht aus 18 unabhängigen Fußballvereinen. Im Jahr 2015 hat sie nach eigenen Angaben einen Gesamtumsatz von 2,6 Mrd. Euro erzielt und verbandsübergreifend über 50.000 Mitarbeiter beschäftigt, was in etwa mit einem MDAX-Unternehmen vergleichbar ist (vgl. Deutsche Fußball Liga GmbH, 2016, S. 2). Während die einzelnen Fußballvereine früher eingetragene Vereine waren, sind sie heutzutage alle in gewinnorientierte Gesellschaften (GmbH oder AG) umgewandelt worden.

Wie in „normalen“ Unternehmen üblich, werden daher auch im Profifußball vermehrt Datenanalysen eingesetzt, um eine bessere Messbarkeit und Steuerbarkeit zu gewährleisten. Dies findet einerseits direkt im sportlichen Teil der Vereine statt, wenn es beispielsweise darum geht, eine Vergleichbarkeit von einzelnen Spielern zu erreichen oder Schwächen zu eliminieren. Andererseits wird Big Data jedoch auch im operativen Teil von Vereinen eingesetzt, wobei dies mit den Ansätzen in normalen Unternehmen vergleichbar ist, da beispielsweise Besucherzahlen von Spielen oder Verkaufszahlen von Merchandising-Produkten zur besseren Planbarkeit vorab geschätzt werden.

Ein recht bekanntes Beispiel für die Verwendung von Statistiken im Fußball ist das Viertelfinale der Weltmeisterschaft 2006 der deutschen Nationalmannschaft, bei dem sich der deutsche Torwart Jens Lehmann vor dem Elfmeterschießen einen Zettel mit den bevorzugten Schusspositionen der gegnerischen Elfmeterschützen geben ließ; Lehmann hielt daraufhin zwei Elfmeter, bei denen die Schützen ihren üblichen Präferenzen entsprechend geschossen hatten (vgl. Steinbrecher und Schumann, 2015, S. 197).

Bei der Weltmeisterschaft 2014 hat die deutsche Nationalmannschaft ebenfalls Statistiken verwendet: der *Deutsche Fußball Bund* (DFB) hat hierfür zusammen mit dem Unternehmen SAP die Smartphone-Anwendung *SAP Match Insights* entwickelt (vgl. SAP SE, 2014). Mit dieser Anwendung konnten die Spieler und Trainer sowohl die eigenen als auch die Spiele der gegnerischen Mannschaften analysieren; diese Analysen wurden einerseits in Form von Statistiken dargestellt, andererseits wurden auch Videoanalysen der gegnerischen Mannschaften integriert.

Ziel der Arbeit

Die dargestellten Beispiele der deutschen Nationalmannschaft lassen sich dem sportlichen Bereich zuordnen, da die Spieler hierdurch ihre eigenen Leistungen und Gegenspieler besser beurteilen können. Da diese Arbeit dem Themenbereich Wirtschaftsinformatik zuzuordnen ist, liegt der Schwerpunkt auf der technischen Unterstützung von Entscheidungsprozessen, d. h. auf betriebswirtschaftlichen und technischen Details, während das Themengebiet der Sportwissenschaft nur angeschnitten werden soll.

In der Theorie müsste beim Aufeinandertreffen zweier Mannschaften immer die bessere Mannschaft gewinnen, was jedoch aufgrund von unvorhersehbaren Ereignissen nicht der Fall ist. Im Rahmen dieser Arbeit soll daher außerdem festgestellt werden, ob durch den Einsatz von Data Mining im Fußball Vorteile realisiert werden können, oder ob nicht der Einfluss des Zufalls oder anderer Faktoren zu groß ist, als dass sinnvolle Ergebnisse erzeugt werden können.

Des Weiteren wird im Rahmen dieser Arbeit eine Analyse auf Datenbeständen echter Fußballspiele durchgeführt, indem versucht wird, das Ergebnis von Fußballspielen anhand von historischen Daten vorherzusagen. Hierzu müssen im ersten Schritt mögliche Kennzahlen und Datenquellen analysiert werden, anhand welcher dann eine Datenbasis aufgebaut werden kann. Auf dieser Basis können anschließend Data-Mining-Algorithmen ausgeführt werden, um eine Prognose von Spielergebnissen zu erhalten.

Aufbau der Arbeit

Im ersten Schritt werden die fachlichen Grundlagen von Data Mining dargestellt, da die Durchführung der Analysen auf diesen basiert. Im Anschluss daran werden die Einsatzmöglichkeiten von Data Mining im Sport beschrieben, wobei zuerst die Ursprünge im Baseball dargestellt werden, um im Anschluss detailliert auf die Einsatzmöglichkeiten innerhalb von Fußballmannschaften einzugehen.

Im nächsten Schritt wird eine Übersicht über mögliche Kennzahlen gegeben, was anhand einer Literaturrecherche veröffentlichter Forschungsergebnisse und verfügbarer Datenquellen durchgeführt wird; hierbei wird zudem dargestellt, durch welche Techniken diese Daten erfasst werden können. Im Anschluss daran wird Data Mining zur Prognose von Spielergebnissen angewandt, wobei hier auch dargestellt wird, wie die Datenbasis anhand der erfassbaren Kennzahlen aufgebaut wird und wie die konkrete Vorgehensweise der Analysen ist. Neben den Analyseergebnissen werden hierbei auch die theoretischen Hintergründe der verwendeten Algorithmen erläutert.

Zum Abschluss werden die Ergebnisse der Arbeit zusammengefasst und es wird ein Fazit gezogen; außerdem wird versucht, einen Ausblick darzustellen, wie die im Rahmen dieser Arbeit durchgeführten Analysen erweitert werden könnten, um die Ergebnisse zu verbessern.

2 Die Grundlagen des Data Mining

In diesem Kapitel werden die Grundlagen des Data Mining dargestellt. Dazu ist es zunächst notwendig, diesen Begriff in einem ersten Schritt von anderen Begriffen abzugrenzen, da im Umfeld von Datenanalysen heutzutage die unterschiedlichsten Begriffe verwendet werden. Im Anschluss daran wird dann der Data-Mining-Prozess (*knowledge discovery in databases*) vorgestellt, welcher das grundsätzliche Vorgehen bei einer Datenanalyse abbildet; außerdem wird dargestellt, wie sich das eigentliche Data Mining in diesen Prozess eingliedert. Zum Abschluss werden die Aufgaben und Methoden des Data Mining näher erklärt; auch wird der Zusammenhang zwischen Aufgaben und Methoden erläutert.

2.1 Begriffsklärung

Als *Data Mining* bezeichnet man die Anwendung von Algorithmen zur Extraktion von neuem Wissen aus großen Datenmengen, wobei unter Wissen in diesem Kontext *Muster* und *Zusammenhänge* verstanden werden, die gültig, nicht-trivial, neu, potentiell nützlich und nachvollziehbar sind (vgl. Fayyad u. a., 1996, S. 40 f.). Eine Definition des Begriffs Data Mining von Grund (2013, S. 31) lautet wie folgt:

> Als Data Mining bezeichnet man Verfahren zur Analyse großer Datenbestände, die aus der unübersehbaren Fülle von Details bisher unbekannte Strukturen und Zusammenhänge heraus filtern und diese Informationen so aufbereiten und bewerten, dass sie eine verständliche Entscheidungshilfe darstellen.

Data Mining wird hierbei im Rahmen des sog. *„Knowledge Discovery in Databases“*-Prozesses durchgeführt, der neben der letztendlichen Durchführung von Algorithmen auch vorbereitende Schritte wie z. B. Datenbereinigung und -transformation beinhaltet (vgl. Abschnitt 2.2). Das Forschungsfeld des Data Mining verdankt seinen Aufschwung der Tatsache, dass aus Unternehmenssicht viele verschiedene Datenquellen vorhanden sind – so gibt es neben internen Datenquellen aufgrund des *World Wide Web* zahlreiche neue Möglichkeiten für die Erfassung von Daten. Das Problem dabei ist nur, dass es nahezu unmöglich ist, manuell Zusammenhänge aus diesen Daten zu erfassen, weshalb dieses Wissen für die Unternehmen zunächst einmal unzugänglich ist.

Die Techniken, die heute im Zusammenhang mit Data Mining verwendet werden, entstammen aus unterschiedlichen Forschungsgebieten:

- **Statistik**: Methoden, die im Rahmen des Data Mining eingesetzt werden, haben zumeist einen mathematischen oder statistischen Hintergrund (vgl. Köppen u. a., 2012, S. 255).
- **Datenbanken**: Data Mining benötigt eine strukturierte Datenbasis – dies können beispielsweise Datensätze in Textdateien oder Datenbanken sein (vgl. Cleve und

Lämmel, 2014, S. 38). Da oftmals große Datenmengen verarbeitet werden müssen, werden Cleve und Lämmel zufolge oftmals relationale Datenbanken gegenüber Textdateien bevorzugt.

- **Visualisierungen**: Die Ergebnisse der Datenanalysen müssen abschließend verständlich dargestellt werden, da die Ergebnisse meistens nicht am Data-Mining-Prozess beteiligte Personen als Adressaten haben; jedoch können manche Methoden (z. B. Clusteranalysen) bereits graphische Modelle als Ergebnis liefern (vgl. Cleve und Lämmel, 2014, S. 14).
- **Künstliche Intelligenz**: Auch zwischen Data Mining und dem maschinellen Lernen aus der künstlichen Intelligenz sind große Überschneidungen vorhanden – in beiden Bereichen geht es um die Analyse von Daten und das Auffinden von Mustern und Zusammenhängen, weshalb dort oftmals Algorithmen verwendet werden, die ihren Ursprung im maschinellen Lernen haben (vgl. Cleve und Lämmel, 2014, S. 13).

Heutzutage werden viele verschiedene Begriffe im Umfeld von Datenanalysen verwendet – der wohl bekannteste hiervon ist der oft als Modewort titulierte Begriff *Big Data*. Doch all diese Begriffe haben grundsätzlich dasselbe Ziel: aufgrund von sehr großen Datenmengen, die im Rahmen des operativen Geschäftes erfasst werden und vom Menschen nicht unmittelbar auswertbar sind, bleiben interessante und evtl. hilfreiche Aussagen und Zusammenhänge innerhalb dieser Daten unentdeckt. Im Zuge von Datenanalysen wird deshalb versucht, Hypothesen über diese zu erstellen.

Cleve und Lämmel (2014, S. 3) bezeichnen die Methoden des Data Mining als „Grundtechniken für neuere und komplexere Ansätze, wie [...] Business Intelligence oder auch Big Data“. Im Folgenden soll eine Erläuterung und Differenzierung der geläufigsten Begriffe gegeben werden, die oft im Zusammenhang mit Data Mining genannt werden:

- **Data Warehouse (DWH)**: Ein DWH dient grundsätzlich der langfristigen Speicherung von Unternehmensdaten, die unterschiedlichsten Quellsystemen entstammen. Eine in der Literatur oft herangezogene Definition eines DWH ist die von Inmon (2005, S. 31):

  > A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions.

  Hieraus wird ersichtlich, dass es darum geht, große und historisierte Datenmengen in einer konsistenten und integrierten Form persistent und effizient zu speichern und die Speicherungsstrukturen für einen schnellen Zugriff zu optimieren; primär spielt hierbei also eine Rolle

  - die Modellierung der Datenstrukturen eines DWH,
  - der Prozess der Datenbeschaffung in Form von Extraktion, Transformation und Laden von Daten in diese integrierte Datenbank und
  - das Ermöglichen anschließender Datenanalysen (vgl. Köppen u. a., 2012, S. 9).

Ein DWH wird in Unternehmen oft zusammen mit Datenanalysen (z. B. Data Mining oder *Online Analytical Processing*, OLAP) eingesetzt – somit findet im DWH eine zentralisierte Bereitstellung der Daten über Geschäftsobjekte statt, die sich sehr gut als Datenbasis für weiterführende Datenanalysen eignet (vgl. Köppen u. a., 2012, S. 8 f.). Eine zwingende Voraussetzung für Data Mining ist ein DWH nach Köppen u. a. jedoch nicht, im Fall von großen, aus verschiedenen Quellen stammenden Daten kann ein DWH jedoch Analysen erleichtern. Der Unterschied vom DWH zum Data Mining liegt also darin, dass es beim Einsatz eines DWH um die optimale Speicherung und Anfragebearbeitung geht, während es beim Data Mining darum geht, durch die Anwendung geeigneter Algorithmen eine Analyse dieser Daten durchzuführen.

- **Business Intelligence (BI)**: Der Ansatz der Business Intelligence ist oberhalb des DWH anzuordnen; hierzu folgende Definition von Kemper u. a. (2010, S. 9):

  > Business Intelligence (BI) bezeichnet einen integrierten, unternehmensspezifischen, IT-basierten Gesamtansatz zur betrieblichen Entscheidungsunterstützung.

  Cleve und Lämmel (2014, S. 3) zufolge sind die Aufgaben von Business Intelligence *Wissensgewinnung*, *Wissensverwaltung* und *Wissensverarbeitung*. Hierbei geht es also nicht nur um das Generieren von Wissen, sondern auch um die Integration von IT-Systemen als Unterstützung in den Bereichen Planung, Steuerung und Entscheidung sowie das damit verbundene Erzeugen und Umsetzen von Strategien zur effektiven Nutzung des generierten Wissens (vgl. Köppen u. a., 2012, S. 250 f.). Daher sind neben Data Warehousing und Data Mining auch andere Tools zur Umsetzung eines BI-Ansatzes in Unternehmen notwendig (vgl. Bauer und Günzel, 2009, S. 13 f.).
- **Big Data**: Im Rahmen des Einsatzes von DWHs können sehr große Datenmengen als Ausgangsgrundlage vorhanden sein; Big Data wird in der Literatur oft durch die drei Eigenschaften *Volume*, *Variety* und *Velocity* definiert, also bezüglich der großen Datenmengen, der Verschiedenheit der Daten und der hohen Geschwindigkeit, in der Daten produziert werden (vgl. Dorschel, 2015, S. 6 ff.). Aus diesen Gründen kann die Datenbasis zu groß und zu komplex sein, als dass klassische Analysemethoden auf Basis bestehender Techniken durchgeführt werden können, weshalb hierbei oftmals innovative Technologien wie z. B. In-Memory- oder NoSQL-Datenbanken eingesetzt werden, die mit eben diesen Datenmengen arbeiten können (vgl. Wu u. a., 2014, S. 6 f.).

Im Rahmen dieser Arbeit liegt der Schwerpunkt auf den klassischen Data-Mining-Methoden, dabei wird an dieser Stelle auch weniger auf die (klassischen) betriebswirtschaftlichen Einsatzmöglichkeiten von Datenanalysen in Unternehmen, wie etwa Business-Intelligence-Ansätze und deren Ziele und Vorteile, eingegangen, sondern mehr auf die technischen Details des Data-Mining-Prozesses und möglicher Algorithmen zur Datenanalyse.

2.2 Der Data-Mining-Prozess

Oft wird Data Mining als Synonym für den ganzen Prozess der Wissenserkennung in Datenbanken (engl. *knowledge discovery in databases*, KDD) verwendet – jedoch ist Data Mining im Grunde genommen nur ein Teilschritt des KDD-Prozesses (vgl. Kemper u. a., 2010, S. 115). Das folgende Modell von Fayyad, Piatetsky-Shapiro und Smyth aus dem Jahr 1996 stellt die Vorgehensweise der Anwendung von Data Mining dar; beginnend mit dem Aufbau der Datengrundlage bis hin zur Interpretation der Ergebnisse.

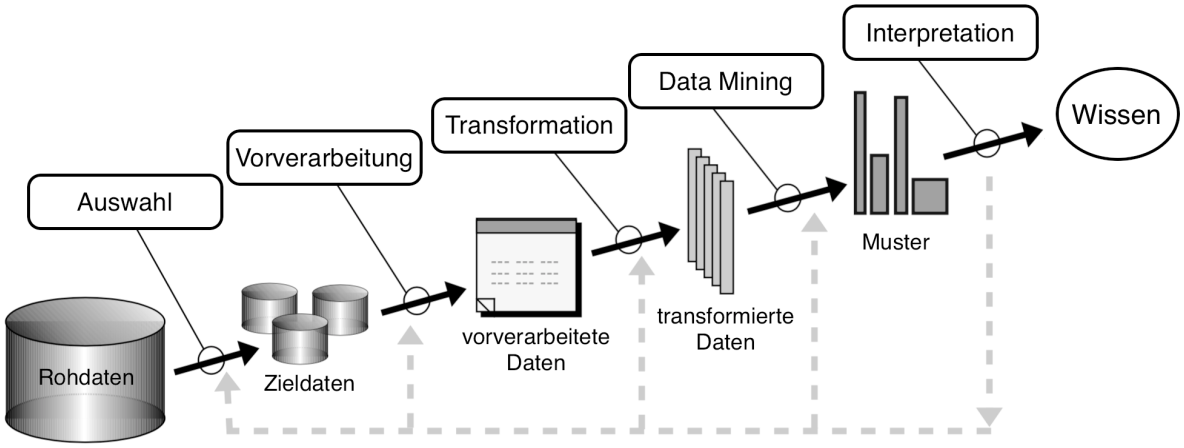


Abbildung 1: Der KDD-Prozess (vgl. Fayyad u. a., 1996, S. 41)

Im Folgenden werden die einzelnen Phasen des KDD-Prozesses aus Abbildung 1 kurz erläutert, angelehnt an die Ausführungen von Fayyad u. a., 1996, S. 42 f. sowie Cleve und Lämmel, 2014, S. 6 f.:

- **Zielidentifikation:** Im Vorfeld der Anwendung müssen die durch das Data Mining zu erreichenden Ziele definiert werden, wozu zwingend domänenspezifisches Wissen notwendig ist, da die Ziele vom jeweiligen Einsatzgebiet abhängen.
- **Auswahl:** Im nächsten Schritt muss die Datengrundlage erzeugt werden, was durch die Auswahl von relevanten Teilen aus einer Menge von Rohdaten stattfindet. Diese Auswahl kann sich sowohl auf die zu betrachtenden Objekte als auch auf deren Merkmale beziehen.
- **Vorverarbeitung:** Im dritten Schritt findet die Bereinigung und Vorverarbeitung der ausgewählten Zieldaten statt; hierbei werden sowohl fehlende als auch fehlerhafte Datenfelder behandelt. Beispielsweise können fehlende Attributwerte durch das Einfügen von Vorgabe- oder Mittelwerten ähnlicher Datensätze ersetzt werden.
- **Datentransformation:** Anschließend erfolgt eine Vorbereitung der Daten für die nachfolgende Analyse. So können hierbei neben der Umwandlung von Attributwerten (z. B. stetige → diskrete Werte) auch Summierungen und Aggregationen durchgeführt oder neue Attribute errechnet werden.
- **Data Mining:** Auf dem für die Analyse vorbereiteten Datenbestand können nun die Data-Mining-Algorithmen durchgeführt werden, die anhand der im ersten Schritt definierten Ziele ausgewählt werden.

- **Interpretation:** Im Anschluss an die Analyse können die erzeugten Daten bzw. Modelle in eine visuelle Form gebracht werden und anschließend interpretiert und mit den definierten Zielen verglichen werden. Je nach dem Ergebnis dieser Interpretationen kann hierbei ein Rücksprung zu den vorherigen Phasen stattfinden, um erneut Analysen durchzuführen.
- **Ergreifung von Maßnahmen:** Zum Abschluss sollte das generierte Wissen den Ansprüchen genügen, somit kann dies an die entsprechenden Fachbereiche zu Ergreifung von Maßnahmen weitergegeben werden oder einfach nur für die Stakeholder dokumentiert werden (vgl. Sharafi, 2013, S. 64).

In der Literatur wird darüber hinaus oftmals auf das sog. CRISP-DM-Referenzmodell verwiesen (siehe Abbildung 2). Dieses Modell, dessen Abkürzung für den *Cross industry standard process for data mining* steht, wurde 1996 in Zusammenarbeit mehrerer Unternehmen (u. a. Daimler AG und SPSS Inc.) mit dem Ziel der Schaffung eines Industriestandards entworfen und seitdem weiterentwickelt (vgl. Chapman u. a., 1999, S. 1). Im Grunde genommen ist es dem KDD-Prozess von Fayyad sehr ähnlich, jedoch steht Chapman u. a. zufolge beim CRISP-DM-Modell der Zyklus der *iterativen* Wissensfindung im Vordergrund. Die im Schaubild abgebildeten Pfeile deuten hierbei lediglich an, was die häufigsten Sprungpunkte sind; der Prozess hat grundsätzlich keinen streng linearen Ablauf, da es gewollt ist, dass neu gewonnene Erkenntnisse zu Rücksprüngen zu vorherigen Phasen führen (vgl. Dorschel, 2015, S. 67 ff.).

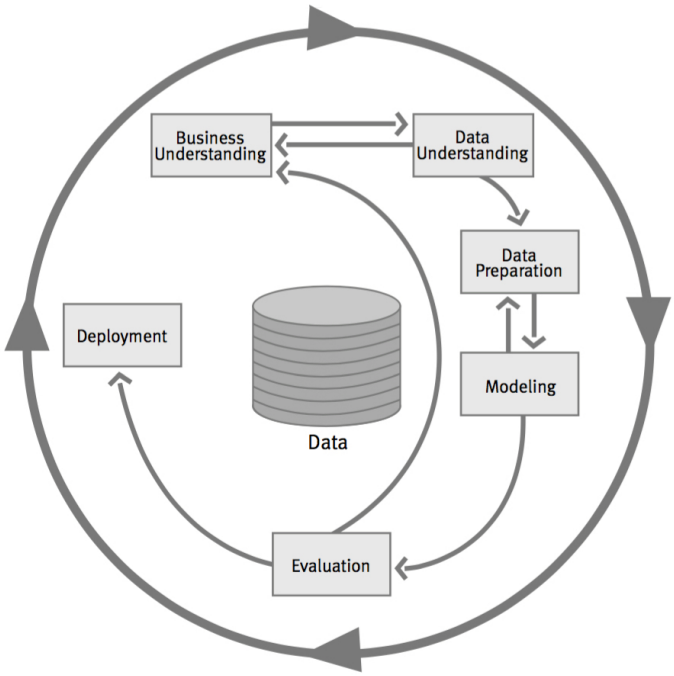


Abbildung 2: Das CRIP-DM-Referenzmodell (Chapman u. a., 1999, S. 10)

Im Bezug auf die einzelnen Phasen sind sich die beiden Vorgehensmodelle sehr ähnlich, das CRISP-DM-Modell orientiert sich lediglich näher an betriebswirtschaftlichen Eigenheiten, während der KDD-Prozess eher technikorientiert ist. Insbesondere im Bezug auf die erste

Phase *business understanding* ist das CRISP-DM-Modell deutlich detaillierter als andere Data-Mining-Vorgehensmodelle (vgl. Sharma und Osei-Bryson, 2009, S. 2114). Aufgrund dessen, dass sich die beiden Modelle grundsätzlich sehr ähnlich sind und dass diese Arbeit vor allem einen Blick auf die technischen Fragen wirft, wird an dieser Stelle aber von einer detaillierteren Erläuterung des CRISP-DM-Modells abgesehen.

2.3 Aufgaben und Methoden des Data Mining

In der Literatur werden die Methoden des Data Mining nach verschiedenen Gesichtspunkten klassifiziert, wobei in diesem Zusammenhang eine Aufteilung auf unterschiedlich viele Ebenen möglich ist; hierbei ist eine Unterteilung auf bis zu drei Ebenen üblich.

So beschränken sich Kemper u. a. (2010) auf eine Ebene, indem sie nur *Aufgaben* darstellen (u. a. Segmentierung und Klassifikation); im Gegensatz hierzu benennt Chamoni (1999) lediglich *Methoden* (u. a. Clusteranalyse und künstliche neuronale Netze). Alpar und Niedereichholz (2000), Bauer und Günzel (2009) und Cleve und Lämmel (2014) schließlich führen eine Unterteilung in zwei Ebenen an: Aufgaben und Methoden, wobei dies einfach eine Darstellung beider einstufigen Unterteilungen ist. Bei Fayyad u. a. (1996) erfolgt sogar eine Unterteilung auf drei Ebenen: hierbei werden auf der ersten Ebene verschiedene *Zielsetzungen* (prediction vs. description) dargestellt, gefolgt von den beiden bereits genannten Ebenen der Aufgaben und Methoden. Diese Zielsetzungen finden sich in den meisten anderen Quellen auch wieder, jedoch wird dies oftmals nicht als eigene Ebene definiert, sondern lediglich bei der Erläuterung der Methoden dargestellt.

Im Folgenden wird Data Mining anhand der von den meisten Autoren verwendeten Unterteilung in Aufgaben (Abschnitt 2.3.1) und Methoden (Abschnitt 2.3.2) unterteilt, wobei im Rahmen der Erläuterung der Aufgaben auch die unterschiedlichen Zielsetzungen dargestellt werden.

2.3.1 Aufgaben des Data Mining

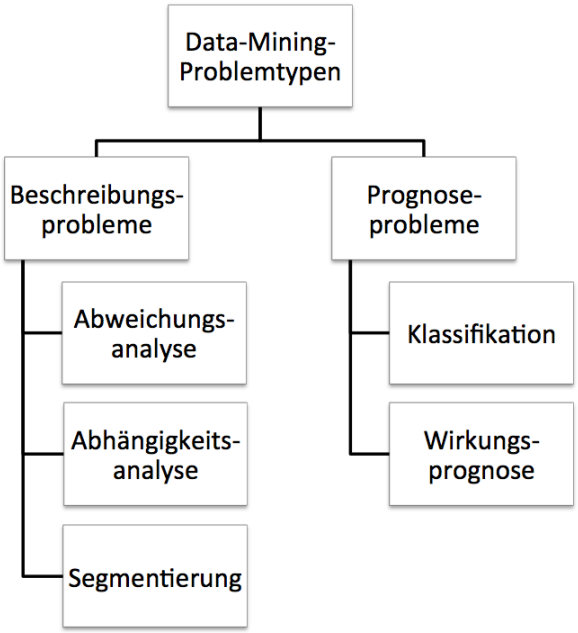


Abbildung 3: Problemtypen im Data Mining (in Anlehnung an Kemper u. a., 2010, S. 115)

Grundsätzlich gibt es zwei verschiedene Problemtypen: Beschreibungsprobleme (*descriptive*) und Prognoseprobleme (*predictive*). Bei Beschreibungsproblemen wird versucht, handlungsrelevante Strukturen innerhalb der Daten zu finden, bei Prognoseproblemen wird hingegen versucht, aus den Werten bekannter Attribute unbekannte oder zukünftige Werte abzuleiten (vgl. Hippner und Wilde, 2009, S. 212). Im Folgenden soll die in Abbildung 3 dargestellte weitere Unterteilung dieser beiden Problemtypen erläutert werden (in Anlehnung an die Ausführungen von Alpar und Niedereichholz, 2000, S. 9 ff. und Vossen, 2008, S. 527 ff.):

- **Abhängigkeitsanalyse:** Hierbei werden Beziehungen innerhalb des Datenbestandes gesucht, wobei dies einerseits Beziehungen zwischen Objekten und andererseits auch Beziehungen zwischen einzelnen Attributen eines Objektes sein können. Als Beispiel hierfür kann die Analyse des Zusammenhangs verschiedener Aktienkurse genannt werden; ein weiteres, oftmals genanntes Beispiel ist die Warenkorbanalyse, bei welcher Artikel gesucht werden, die häufig zusammen gekauft werden.
- **Abweichungsanalyse:** Im Gegensatz hierzu werden bei der Abweichungsanalyse Objekte gesucht, die eine größtmögliche Abweichung zu den restlichen Datensätzen haben. Algorithmen, die diese Aufgabe unterstützen, könne einerseits komplementär zu anderen Algorithmen angewandt werden – beispielsweise zur Vorbereitung für eine folgende Datenanalyse, um fehlerhafte Datensätze oder Ausreißer zu finden und zu eliminieren; andererseits finden sich derartige Aufgabenstellungen beispielsweise im Bankensektor wieder, wo auffällige Transaktionen zum Auffinden von Betrugsfällen identifiziert werden (vgl. Kemper u. a., 2010, S. 116).

- **Segmentierung:** Bei der Segmentierung geht es darum, eine Menge an Datensätzen in disjunkte Teilmengen, sog. Cluster zu unterteilen, wobei die einem Cluster angehörigen Objekte möglichst homogene Attributwerte haben sollten. Die verschiedenen Teilmengen werden hierbei nicht im Vorhinein definiert, sondern erst im Rahmen der Analyse generiert, weshalb entsprechende Algorithmen die Attribute, die diese Cluster definieren, selbständig identifizieren müssen. Als Beispiel hierfür kann das Auffinden von Kundengruppen genannt werden, wobei die Unterscheidung zwischen diesen z. B. anhand der Zahlungsweise oder dem Kaufzeitpunkt stattfinden kann.
- **Klassifikation:** Im Falle der Klassifikation werden die betrachteten Datensätze vorher definierten Klassen zugeordnet, wobei das Ziel ist, dass deren Merkmale eine möglichst große Übereinstimmung mit den charakteristischen Merkmalen der Klassen besitzen. Da anders als bei der Segmentierung die Klassen im Vorhinein definiert sind, müssen die Algorithmen lediglich feststellen, welcher Klasse ein Objekt angehört und nicht, wie diese Klassen definiert werden könnten. Als Beispiel kann hier die Analyse von Bildern von Fahrzeugen angeführt werden, wobei die Zuordnung bspw. zu den Klassen Auto, LKW und Motorrad stattfinden kann.
- **Wirkungsprognose:** Bei der Wirkungsprognose werden unbekannte Attributwerte vorhergesagt, wobei der Vorhersagewert einerseits durch die Analyse früherer Werte des zu prognostizierenden Attributes (Zeitreihen) errechnet werden kann; andererseits kann dieser jedoch auch anhand anderer Attribute des Objektes generiert werden. In den meisten Fällen besitzt das zu ermittelnde Attribut einen Zukunftsbezug; im Gegensatz zur Klassifikation erfolgt hierbei i. d. R. auch die Feststellung eines numerischen Wertes, beispielsweise bei der Prognostizierung des zukünftigen Auftragsvolumen anhand einer Kaufhistorie (vgl. Kemper u. a., 2010, S. 117).

In der Literatur werden die Bereiche Text Mining und Web Mining teilweise als selbständige Aufgaben neben Data Mining genannt (u. a. bei Cleve und Lämmel, 2014 und Kemper u. a., 2010 sowie Vossen, 2008), eigentlich sind dies jedoch nur Spezialisierungen von Data Mining, die auf Methoden und Algorithmen des Data Mining zurückgreifen. Der Vollständigkeit halber sei im Folgenden eine kurze Erläuterung in Anlehnung an Kemper u. a. (2010, S. 117 ff.) gegeben:

- **Text Mining:** Im Gegensatz zum klassischen Data Mining, wo die Datenbasis strukturiert und somit direkt maschinenverarbeitbar ist, werden beim Text Mining unstrukturierte Daten analysiert, was beispielsweise Kundenbewertungen, Emails oder andere Freitexte sein können. Da die zu analysierenden Daten unstrukturiert sind, ist die größte Herausforderung hierbei, dass derartige Systeme und Algorithmen gleichsam ein gutes Textverständnis benötigen, um die Texte interpretieren zu können.
- **Web Mining:** Im Gegensatz zum Text Mining basiert das Web Mining auf einer semistrukurierten Datengrundlage, was im Rahmen des *World Wide Web* neben Logdateien auch Webseiten sein können, da diese neben Text-Elementen auch eine

gewisse Strukturierung vorweisen. Ein beispielhaftes Anwendungsszenario für das Web Mining lässt sich innerhalb des Marketings finden, wo die Extraktion von Meinungen über Produkte aus Bewertungsportalen durchgeführt werden kann.

2.3.2 Methoden des Data Mining

Zur Bearbeitung der unterschiedlichen Aufgabenstellungen des Data Mining gibt es viele verschiedene Methoden, deren Ursprung – wie bereits gesagt – oftmals in der Statistik liegt. Die geläufigsten Methoden werden in der Abbildung 4 dargestellt, die Pfeile deuten hierbei an, welche Aufgabenstellungen die jeweiligen Methoden verwenden können.

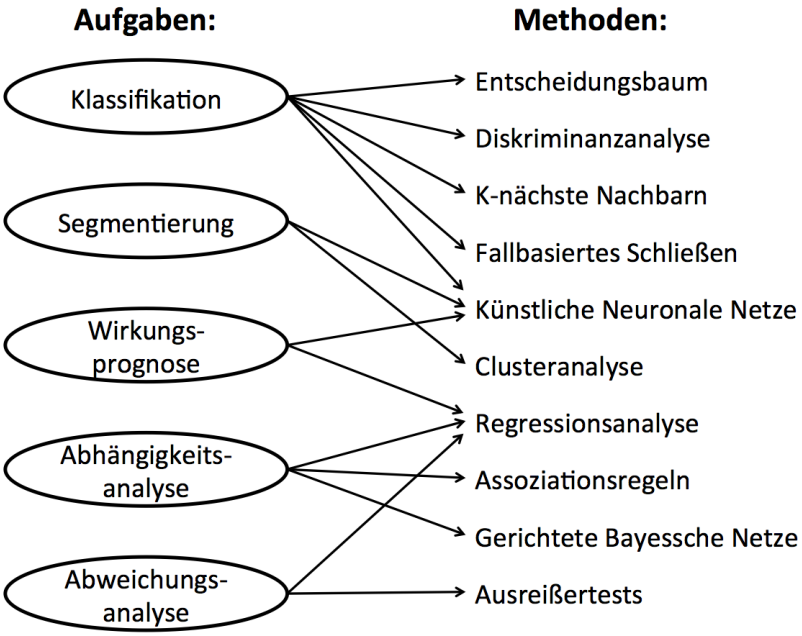


Abbildung 4: Zuordnung von Data-Mining-Methoden zu -Aufgaben (in Anlehnung an Alpar und Niedereichholz, 2000, S. 13)

Im Kontext dieser Arbeit werden verschiedene Methoden eingesetzt; eine Erläuterung dieser erfolgt an den Stellen, an denen sie eingesetzt werden (Abschnitte 5.7.2 bis 5.7.5).

2.4 Grenzen und Probleme

Beim Einsatz von Data-Mining-Technologien gibt es verschiedene Problembereiche; durch strukturiertes Vorgehen (z. B. KDD-Prozess) wird jedoch versucht, diesen Einhalt zu gebieten. Im Folgenden werden grundsätzliche Problembereiche angeschnitten; diese Aufzählung ist jedoch bei weitem nicht vollständig, sondern soll nur einen Anhaltspunkt bieten.

Neben der grundsätzlichen Problematik bezüglich der Qualität der Datengrundlage, was im Rahmen des KDD-Prozesses zu lösen versucht wird (z. B. fehlende/falsche Werte, Ausreißer etc.) liegt ein weiteres Problem darin, dass der Umfang der zugrundeliegenden Daten für Data Mining immer größer wird. Während es früher um eine Anwendung klassischer

Methoden der Statistik ging, müssen Data-Mining-Anwendungen heutzutage auf die geänderten Anforderungen reagieren. Die großen Datenmengen haben einerseits Auswirkungen auf die verschiedenen Algorithmen des Data Mining, die diese Datenmengen verarbeiten müssen; anderseits hat dies auch Auswirkungen auf die Speicherung der Daten, wo daher beispielsweise optimierte Data Warehouses als Ersatz für normale Datenbanken oder Textdateien eingesetzt werden.

Algorithmen, die im Zuge des Data Minings eingesetzt werden, benötigen eine Parametrisierung – als Beispiel hierfür kann der sog. *DBSCAN*-Algorithmus genannt werden, welcher zur Clusteranalyse verwendet werden kann; dieser benötigt zwei Parameter, die minimale Punkteanzahl für Cluster und den maximalen Abstand zwischen zwei Punkten (Umgebungsradius) innerhalb eines Clusters (vgl. Cleve und Lämmel, 2014, S. 160 ff.). Werden Cleve und Lämmel zufolge für diese beiden Parameter nun sehr geringe Werte ausgewählt, erzeugt der Algorithmus deutlich mehr Cluster als sinnvoll (wenn das Minimum auf 1 gesetzt wird, gäbe es im Extremfall sogar Cluster mit nur einem Objekt); daher muss dies bei der Anwendung der verschiedenen Algorithmen berücksichtigt werden, ggf. muss mit diesen Werten experimentiert werden.

Die erarbeiteten Ergebnisse müssen im Anschluss an das Data Mining validiert und interpretiert werden, da die Methoden lediglich ihre Algorithmen ausführen und weder einschätzen können, inwiefern die Ergebnisse sinnvoll und nicht-trivial sind, noch haben sie Hintergrundwissen über die Bedeutung der Kennzahlen (vgl. Knobloch und Weidner, 2000, S. 6 ff.). Im betrieblichen Umfeld ist es logisch, dass die Ergebnissen zu verwerten sind, da sich die Kosten des Einsatzes von Data Mining amortisieren müssen. Um etwaige Handlungsmaßnahmen zu generieren sind neben analytisch-methodischen auch fachliche Kenntnisse notwendig, da dies ein nicht-trivialer Prozess ist; ein Beispiel hierfür sind Clusteranalysen, die den Datenbestand in verschiedene Cluster unterteilen, jedoch keine Aussagen über die Bedeutungen der gefundenen Cluster treffen können (vgl. Knobloch und Weidner, 2000, S. 6 ff.).

3 Data Mining im Fußball

Heutzutage wird Data Mining in vielen professionell ausgeführten Sportarten eingesetzt; neben Fußball lassen sich hier unter anderem Golf, Schwimmen, Basketball, Radrennen, Rugby, Rudern, Leichtathletik, American Football und Tischtennis nennen (vgl. Ofoghi u. a., 2013, S. 173). Allerdings eignet sich nicht jede Sportart für den Einsatz von Data Mining, wobei dies Ofoghi u. a. zufolge von der grundsätzlichen Struktur des Spiels abhängig ist.

Im Rahmen dieser Arbeit wird grundsätzlich nur der Anwendungsfall Fußball betrachtet; in Abschnitt 3.1.2 wird aber zusätzlich auf den Einsatz im Baseball eingegangen, da dort die Ursprünge für den Einsatz von Data Mining im Sport liegen. Bei der Darstellung des Fußballs werden einerseits die sportwissenschaftliche Theorie, die hinter dem Einsatz von computergestützten Systemen steht, und andererseits die konkreten Einsatzmöglichkeiten von Data Mining im und um den Fußball behandelt. Zum Abschluss erfolgt eine kritische Betrachtung durch die Erläuterung verschiedener Problembereiche.

3.1 Sport Data Mining

Küppers (1998, S. 148 f.) zufolge eignet sich Data Mining insbesondere für Bereiche, in denen kleine Verbesserungen von zentraler Bedeutung für den Geschäftserfolg sein können – was für den Einsatz im Sport zutrifft. Bei den Individualsportarten kann eine sehr kleine Differenz schon Sieg oder Niederlage bedeuten, so z. B. im Motorsport, wo es um Millisekunden geht. Im Mannschaftssport ist dies jedoch nicht anders, der Unterschied liegt darin, dass es hierbei nicht um ein konkretes Ergebnis geht, sondern um die Mannschaftsleistung, die auf ganz verschiedenen Parametern basiert, beispielsweise kann die Auswahl von Spielern für eine Mannschaft Einfluss auf das Spielergebnis haben.

Betrachtet man den Sport nun betriebswirtschaftlich, so kann der Unterschied zwischen Sieg und Niederlage große finanzielle Auswirkungen haben. Als Beispiel hierfür kann die im Rahmen dieser Arbeit betrachtete Sportart Fußball herangezogen werden: der Abstieg aus der ersten deutschen Bundesliga in die zweite Liga führt beispielsweise zu Umsatzeinbrüchen von durchschnittlich 33 % (vgl. Laub und Merx, 2013), des Weiteren werden die zu verteilenden Vermarktungseinnahmen von mehreren hundert Millionen Euro anhand von Verteilungsschlüsseln errechnet, denen u. a. die Endplatzierungen in der Tabelle zugrunde liegen – eine Verbesserung um eine Position in der Abschlusstabelle führt zu Mehreinnahmen in Millionenhöhe (vgl. Deutsche Fußball Liga GmbH, 2012).

3.1.1 Stufenmodell der Verwendung von Data Mining im Sport

Im Sport werden sehr viele Daten erfasst – während es früher nur die wichtigsten Ereignisse waren (z. B. für Zeitungsberichte) werden heutzutage sehr große Datenmengen erzeugt, die von den Mitarbeitern von Sportmannschaften analysiert werden können, um

Schlüsse aus diesen zu ziehen und entsprechende Maßnahmen zu ergreifen. Aufgrund der unüberschaubaren Datenmengen ist es heutzutage nicht mehr bzw. nur mit sehr großem Aufwand möglich, diese manuell zu analysieren – daher müssen hier geeignete Vorgehensweisen zur automatisierten Durchführung dieser Analysen eingesetzt werden. Jedoch wird nicht in allen Sportarten gleich auf diese Daten zurückgegriffen, in Tabelle 1 wird ein Stufenmodell des Einsatzes von Daten im Sport von Schumaker u. a. (2010) dargestellt.

Stufe	**Art der Verwendung von Sportdaten**
1	Keine Verwendung von Sportdaten
2	Vorhersagen durch Experten anhand von Erfahrungswerten
3	Vorhersagen durch Experten unter Verwendung historischer Daten
4	Verwendung von Statistiken im Entscheidungsprozess
5	Verwendung von Data Mining im Entscheidungsprozess

Tabelle 1: Stufenmodell der Verwendung von Sportdaten (vgl. Schumaker u. a., 2010, S. 2)

Die Bedeutungen der einzelnen Ebenen werden im Folgenden erläutert, angelehnt an die Ausführungen von Schumaker u. a. (2010, S. 2 f.):

Auf der ersten Ebene findet überhaupt keine Verwendung von Statistiken statt, wie es bspw. im Amateursport der Fall ist, da der Spaß hierbei im Vordergrund steht. Daten werden hier entweder überhaupt nicht oder nur zu historischen Zwecken aufgezeichnet.

Auf der zweiten Ebene greifen Domainexperten auf ihre Erfahrungswerte und ihr „Bauchgefühl“ zurück; konkrete Daten oder Statistiken finden hierbei jedoch keine Verwendung. Als Beispiel hierfür werden von Schumaker u. a. Trainer genannt, die Auswechslungen durchführen, weil sich diese Entscheidungen „richtig“ anfühlen. Im Gegensatz hierzu findet auf der dritten Ebene eine Verwendung historischer Daten statt, indem die Trainer beispielsweise Strategien anwenden, die sich in der Vergangenheit als erfolgreich herausgestellt haben.

Auf der vierten Ebene werden die Domainexperten im Entscheidungsprozess durch Statistiken unterstützt, wobei dies Schumaker u. a. zufolge sowohl einfachere (z. B. Anzahl von Ballkontakten) als auch Statistiken komplexerer Art sein können. Als Beispiel für letzteres nennen sie die Bewertung der Spieler einer Mannschaft durch das Errechnen spezifischer Kennzahlen, welche die Trainer später berücksichtigen können.

Auf der letzten Ebene basiert der Entscheidungsprozess auf Data Mining, wobei der Unterschied zur vierten Ebene nach Schumaker u. a. darin liegt, dass die Ergebnisse des Data Mining einerseits deutlich komplexer sind und andererseits die Bedeutung der Daten im Entscheidungsprozess deutlich größer sein kann, da die Domainexperten im Extremfall außen vor gelassen werden können, so dass die Entscheidungen unabhängig von menschlichem Input gefällt werden. Die Grenze zwischen der vierten und fünften Ebene ist hierbei fließend, da die Berechnung von Statistiken recht komplex sein und bereits in Richtung Data Mining tendieren kann.

Der Hauptvorteil des Einsatzes von Data Mining lässt sich aus dem folgenden Zitat von Schumaker u. a. (2010, S. 3) recht gut erkennen:

> By removing the potential of human biases from the decision making process, coaches and managers have the potential to manage more effectively and make objective decisions that can help the organization.

Offensichtlich wäre es (ihrer Meinung nach) besser, die *subjektiven Entscheidungskriterien* herauszunehmen, da die Trainer beispielsweise gewisse Spieler bevorzugen können, obwohl diese im Vergleich zu anderen Spielern „schlechter“ sind (z. B. Lieblingsspieler oder wenn Trainer und Spieler verwandt sind). Der Extremfall – also das komplette Eliminieren der Domainexperten aus dem Entscheidungsprozess – ist jedoch recht selten, da es neben quantitativen Kennzahlen auch recht viele nicht maschinell auswertbare Einflüsse gibt (vgl. Schumaker u. a., 2010, S. 3). Beispielsweise könnten bei der Verpflichtung von neuen Spielern zwar das Finden dieser Spieler computergestützt stattfinden, so dass diese optimal zur Strategie passen, jedoch müssen die Spieler auch von ihrer Persönlichkeit her ins Mannschaftsgefüge passen, was sich üblicherweise in persönlichen Gesprächen zwischen Spieler und Trainer herausstellt. Daher ist es Schumaker u. a. zufolge heutzutage üblich, dass sich professionelle Mannschaften auf der vierten oder fünften Ebene befinden, wo Statistiken bzw. Data Mining im Entscheidungsprozess unter Heranziehung der Expertise von Domainexperten zur Hilfe gezogen werden.

3.1.2 Ursprünge

Heutzutage mag es in vielen Sportarten üblich sein, Data-Mining-Methoden einzusetzen; früher war dies jedoch alles andere als üblich. Die Ursprünge des Einsatzes von Data Mining im Sport liegen im Baseball, wo bereits 1970 vereinzelte Trainer intensiv mit Statistiken arbeiteten (vgl. Lewis, 2004, S. 91 f.). Das bekannteste Beispiel für Data Mining im Sport sind die sog. *Moneyball*-Jahre Anfang der 2000er-Jahre im Baseball. Billy Beane, Manager der Mannschaft *Oakland Athletics*, stand 1997 vor folgender Herausforderung (Lewis, 2004, S. 119):

> You have $40 million to spend on twenty-five baseball players. Your opponent has already spent $126 million on its own twenty-five players, and holds perhaps another $100 million in reserve. What do you do with your forty million to avoid humiliating defeat?

Aufgrund dieser Unausgeglichenheit mussten die kleineren Mannschaften mit geringeren Budgets erfinderisch werden und waren daher auf der Suche nach Schnäppchen – d.h. nach jüngeren, älteren, unbekannten oder von Konkurrenten unterschätzen Spielern (vgl. Lewis, 2004, S. 119 f.). Der sog. *Sabermetrics*-Ansatz (angelehnt an das Akronym *SABR, Society for American Baseball Research*), in dessen Rahmen die Anwendung von Statistiken im Baseball erforscht wurde, war bereits seit 1990 in wenigen Teams relativ erfolglos

im Einsatz; Beane versuchte nun, sich auf diesen Ansatz zu fokussieren (vgl. Lewis, 2004, S. 81 ff.). Hierbei werden Korrelationen zwischen erfolgreichen Mannschaften und Kennzahlen erforscht, wobei diese Kennzahlen im Gegensatz zu früheren Versuchen nicht auf einfachen Statistiken beruhten (z. B. dem *batting average*, der durchschnittlichen Anzahl an Schlägen eines Spielers), da diese nur Teilaspekte eines Spielers betrachten (vgl. Grabiner, 1994). Daher findet hierbei Grabiner zufolge eine Entwicklung komplexerer Kennzahlen statt, deren Berechnungsgrundlage verschiedene statistische Werte sind, mit dem Ziel der Bewertung der Gesamtleistung von Spielern.

Im Baseball geht es grundsätzlich darum, die *playoffs* zu erreichen, für die sich die besten Teams der *regular season* qualifizieren. Das Team um Beane hatte durch die Analyse historischer Daten festgestellt, dass es in den meisten Fällen genügte, 95 der insgesamt 162 Spiele zu gewinnen, um diese Endspiele zu erreichen (vgl. Lewis, 2004, S. 124). Um diese Quote zu erreichen, muss Lewis zufolge die Offensive des eigenen Teams, wie die Berechnungen von Beane ergaben, im Durchschnitt mindestens 135 *runs* mehr erzielen als die gegnerische Mannschaft, wobei die eigene Offensive pro Spiel zwischen 800 und 820 runs erreichen sollte, während die Defensive maximal 650 bis 670 runs zulassen sollte.

Im Rahmen des *Sabermetrics*-Ansatzes wurde nun versucht, Kennzahlen für Spieler herauszufinden, die eine starke Korrelation mit Siegen mit einer Differenz von mindestens 135 *runs* bzw. mit den oben genannten offensiven und defensiven Leistungen hatten (vgl. Lewis, 2004, S. 124). In den folgenden Jahren verpflichtete das Team um Beane, Lewis zufolge, neue Spieler daher nicht wie üblich anhand einer subjektiven Beobachtung und Bewertung, sondern komplett anhand dieser Kennzahlen; daher wurden hierbei auch einige Spieler verpflichtet, die von anderen Teams ignoriert wurden, die auf klassische Eigenschaften wie z. B. Athletik achteten. Darüber hinaus wurde anhand der errechneten Bewertungen der eigenen Spieler versucht, Vorhersagen über den Ausgang einer Saison zu erstellen; entsprechend wurde das Team durch die Verpflichtung neuer Spieler verbessert, sollten die Voraussagen ergeben, dass die genannten Ziele nicht zu erreichen wären (vgl. Lewis, 2004, S. 124 f.).

Nach einer Anlaufphase war Beane mit seinem Team erfolgreich, nach acht Jahren ohne *Playoff*-Qualifikation hatten sich die *Oakland Athletics* im Jahr 2000 erstmals qualifiziert, gefolgt von drei weiteren erfolgreichen Jahren bis 2003 (vgl. Oakland Athletics, 2014). Das Team schied aber jeweils in der ersten Runde (vergleichbar mit einem Viertelfinale im Fußball) aus; begründet wird dies damit, dass der Zufall in den *playoffs* einen größeren Einfluss hat als in der *regular season*, da sich dieser in einer Saison mit 162 Spielen eher ausgleicht als in einem Aufeinandertreffen mit nur 5 Spielen (Lewis, 2004, S. 274):

> Over a long season the luck evens out, and the skill shines through. But in a series of three out of five [...] anything can happen. In a five-game series, the worst team in baseball will beat the best about 15 percent of the time [...]

In der Folgezeit war das Team um Beane jedoch nur noch mäßig erfolgreich; die Gründe hierfür liegen vermutlich darin, dass der *Sabermetrics*-Ansatz heutzutage im Baseball üblich ist, weshalb die Wettbewerbsvorteile gegenüber anderem Teams geringer wurden und Teams mit mehr Geld wieder erfolgreicher waren.

3.2 Besonderheiten des Fußballs

Es gibt wesentliche Unterschiede zwischen den einzelnen Sportarten, die dazu führen, dass sich manche Sportarten besser für den Einsatz von Data Mining eignen. Im Folgenden sollen die diesbezüglichen Eigenschaften des Fußballs im Vergleich zu anderem Sportarten erläutert werden.

Im Baseball und im American Football dauern die verschiedenen Spielzüge der Mannschaften beispielsweise nur wenige Sekunden und sind in sich abgeschlossen – es wird z. B. ein Ball durch Mannschaft A geworfen, die gegnerische Mannschaft B reagiert; im Anschluss bekommt je nach Ausgang des abgeschlossenen Spielzugs entweder Mannschaft A oder B den Ballbesitz und darf einen neuen Spielzug durchführen. Im Gegensatz hierzu ist der Fußball etwas komplexer, da der Ballbesitz durch Zweikämpfe jederzeit wechseln kann; außerdem haben die Spielzüge keinen festen Aufbau, sondern sind von der jeweiligen Strategie der Teams abhängig. So gibt es beispielsweise Ballbesitzstrategien, wo der Spielaufbau über viele Pässe durchgeführt wird und somit länger dauert; im Gegensatz hierzu stehen Kontertaktiken, bei welchen sich der Spielaufbau auf möglichst schnelle Gegenangriffe fokussiert.

Darüber hinaus gibt es zahlreiche weitere Eigenheiten, welche die Sportart Fußball ausmachen. Im Folgenden wird die Aufstellung von Czwalina (1988, S. 60) wiedergegeben; diese hat sich im Laufe der Jahre als nahezu vollständig herausgestellt und wird daher oft zum Vergleich von anderen Sportarten mit dem Fußball herangezogen:

- große Fläche des Spielfeldes und damit vermachte große Aktionsradien der Spieler,
- hohe Zahl von Spielern,
- Sonderstellung des Torwarts,
- „zufällige“ Ballwege durch „handloses“ Spiel,
- geringe Zahl von Torerfolgen,
- häufiger körperlicher Kontakt von Gegenspielern und damit verbundene „Gefahr“ von Fouls,
- „arbeitsteiliges“ Spielerverhalten bedingt, dass nicht die gesamte Mannschaft angreift oder verteidigt,
- ausgeprägtes Mittelfeldspiel,
- geringe Zahl von Ballkontakten eines Spielers in einer Zeiteinheit,
- zeitgebundenes (statt satzgebundenes) Spiel von langer Dauer,
- Abhängigkeit von Platz- und Witterungsverhältnissen,
- große Streuung des „Raumstellenwertes von Spielvorgängen“.

Aufgrund dieser Eigenschaften des Fußballs lassen sich die im Baseball und American Football erarbeiteten Modelle nicht 1:1 für den Fußball anwenden; auch mit Anpassungen ist dies nur schwer möglich, da die Sportarten doch sehr verschieden sind. Andererseits gibt es jedoch Sportarten, die sich sehr ähnlich sind; hier bietet sich der Versuch einer Übertragung der Modelle an. Als Beispiel können hier die Spiele Kricket und Baseball genannt werden, die einer gemeinsamen Basis entstammen (vgl. Schumaker u. a., 2010, S. 49). Ein weiteres Beispiel sind die beiden Ballsportarten Hockey und Fußball, deren Unterschied grundsätzlich darin besteht, dass der Ball beim Hockey mit Schlägern, während er im Fußball mit dem Fuß gespielt wird (vgl. Schumaker u. a., 2010, S. 95).

Versuche, die den Modellen übergeordneten, grundsätzlichen Methoden zu übertragen sind jedoch üblich, da diese den Einsatz von Data Mining im Sport auszeichnen; der in Kapitel 3.1.2 verwiesene *Sabermetrics*-Ansatz wurde im Basketball analog durch den sog. *APBRmetrics*-Ansatz umgesetzt (durch die *Association of Professional Basketball Researchers*), die Ziele hierbei waren dieselben wie beim Vorbild im Baseball, nämlich die Erstellung komplexerer, ganzheitlicher Kennzahlen anstatt der Verwendung einfacher Statistiken (vgl. Schumaker u. a., 2010, S.27).

3.3 Anspruchsgruppen

Grundsätzlich befinden sich die primären Interessengruppen des Einsatzes von Data Mining im Fußball natürlich in den Vereinen und es sind insbesondere die Trainer, die ihre betreuten und die gegnerischen Mannschaften zum Zweck der Spielvorbereitung analysieren wollen. Bauer (1998, S. 13) zufolge gibt des neben den Trainern noch die beiden Gruppen der Journalisten und der Sportwissenschaftler. Hierbei greifen Journalisten im Rahmen ihrer Berichterstattungen auf Statistiken zurück, wobei diese Analysen, Bauer zufolge, trotzdem meistens sehr subjektiv geprägt sind. Sportwissenschaftler hingegen versuchen in ihren Analysen auf Basis erfasster Daten durch die Anwendung wissenschaftlicher Methoden neue Erkenntnisse zu gewinnen, als Beispiele hierfür nennt Bauer taktische Verhaltensmuster oder die Weiterentwicklung von Spielsystemen.

Neben diesen Gruppen haben heutzutage die Spieler selbst ein Interesse an den Ergebnissen von Datenanalysen, da sie sich selbst weiterentwickeln und ihre Spielleistungen bewerten wollen. Einige Sportler veröffentlichen Daten über sich auf ihren Social-Media-Auftritten, beispielsweise der deutsche Nationalspieler Toni Kroos, der regelmäßig Statistiken des Anbieters *Opta Sports* (vgl. Kapitel 4.3) auf seiner Website veröffentlicht (siehe Abb. 5).

Abbildung 5: Statistikdaten Website Toni Kroos (vgl. Kroos, 2016)

In der sportwissenschaftlichen Literatur wird in der Regel nur auf die oben genannten Gruppen (Trainer, Spieler, Journalisten und Sportwissenschaftler) verwiesen, jedoch gibt es heutzutage noch einige weitere Interessengruppen. So wecken Statistiken und Analysen das Interesse von Fans von Fußballvereinen, da diese sich für die Leistungen „ihrer" Mannschaften interessieren. Im American Football ist es beispielsweise auf den „besseren" Sitzplätzen üblich, dass dort Tablet-PCs angebracht sind, mit denen die Zuschauer Echtzeit-Analysen und Wiederholungen von Spielszenen in Form von Videos anschauen können (vgl. Steinbrecher und Schumann, 2015, S. 200 f.). Doch auch außerhalb von Spielen finden sich Statistiken im Fokus der Fußballinteressierten, da es für viele Fußballfans üblich ist, an Tippspielen teilzunehmen, bei denen die Vorhersage von Spielergebnissen logischerweise möglichst akkurat sein sollte.

Darüber hinaus interessieren sich auch Unternehmen, die direkt mit dem Fußball in Berührung kommen, für solche Datenanalysen. Als Beispiel hierfür können Wettanbieter genannt werden, deren Geschäftsmodell darauf basiert, dass die Wettquoten auf den Ergebnissen von Datenanalysen basieren (Wahrscheinlichkeiten von Siegen bzw. Niederlagen). Außerdem werden dort seit einigen Jahren *Fraud-Detection*-Systeme eingesetzt, um Spielmanipulationen zu erkennen – hierzu werden von den übergeordneten Verbänden (z. B. DFB, UEFA und FIFA) sämtliche Wettmärkte in Echtzeit analysiert, um auffällig hohe Wettbeträge zu lokalisieren (vgl. UEFA, 2014).

Außerdem haben weitere, mit dem Fußball verbundene Unternehmen, wie z. B. Sponsoren oder Händler ein Interesse an Vorhersagen für den Erfolg von Mannschaften, da diese einen Einfluss z. B. auf Verkaufszahlen haben (bspw. *Merchandising*/Fan-Artikel wie Trikots).

3.4 Einsatzmöglichkeiten

In vielen Vereinsmannschaften wird heutzutage auf Datenanalysen gesetzt, manche Vereine verwenden hier fertige Produkte von Dienstleistern, andere haben eigene Abteilungen, die mit den Datenanalysen betraut sind. Beispielsweise beschäftigt der FC Bayern Mün-

chen eigenen Angaben zufolge 10 Mitarbeiter (vgl. FC Bayern München AG, 2015), der dänische Verein FC Midtjylland kann den Aussagen seines Sportdirektors zufolge auf ein Team aus 200 Datenanalysen zurückgreifen (vgl. Biermann, 2015). Dieser dänische Verein ist jedoch ein Spezialfall – Biermann zufolge war es Matthew Benham, der durch die Anwendung von Data Mining im Bereich der Sportwetten reich wurde und Anteile am FC Midtjylland gekauft hat, um dort Data Mining als zentrales Mittel zum Erfolg zu etablieren. Mangat (2015) zufolge hat Benham hier in vielen Bereichen Kennzahlen eingeführt, beispielsweise wurde der Erfolg dort nicht wie üblich durch die Tabellenposition beurteilt, sondern durch Kennzahlen, die eine starke Korrelation mit langfristigem Erfolg haben.

Üblicherweise wird Data Mining jedoch nicht derart „extrem“ umgesetzt, in den meisten Vereinen wird findet sich Data Mining nur in wenigen Bereichen wieder, da letztlich die Trainer Entscheidungen wie z. B. die Aufstellung in Spielen treffen wollen.

Nun wird erst einmal dargestellt, in welchen Bereichen Fußball mit Data Mining analysiert werden kann, wobei an dieser Stelle nur die Einsatzmöglichkeiten dargestellt werden, auf die Darstellung des Einsatzes konkreter Data-Mining-Verfahren wird verzichtet, da der Schwerpunkt dieser Arbeit auf der Anwendung von Verfahren zur Prognose von Fußballspielen liegt (siehe Kapitel 5).

3.4.1 Vor- und Nachbereitung von Spielen

Lames (1994, S. 21 f.) zufolge lässt sich der gesamte Prozess der Anwendung von Datenanalysen im Rahmen der Spielvor- und -nachbereitung in drei Phasen unterteilen:

- **Beschreibungsphase:** Um Analysen durchzuführen, muss eine Datenbasis erstellt werden, die das Spielgeschehen beschreibt; Lames schlägt hierzu den Prozess der systematischen Spielbeobachtung vor.
- **Diagnosephase:** Aufbauend auf dem Datenbestand können nun „Deutungen, Interpretationen und Analysen des Wettkampfverhaltens“ durchgeführt werden (vgl. Lames, 1994, S. 25).
- **Trainingspraktische Umsetzung:** Im letzten Schritt können die Ergebnisse der Datenverarbeitung im Training umgesetzt werden; hierbei müssen geeignete Trainingsziele entwickelt werden, aus welchen eine trainingsmethodische Auswahl abgeleitet werden kann (vgl. Lames, 1994, S.28).

Basierend auf den Ergebnissen der Spielbeobachtung im Rahmen der Beschreibungsphase kann anschließend durch die Durchführung von Analysen Wissen generiert werden. Diese Analysen können nicht nur zur Spielvor- und -nachbereitung verwendet werden, sondern auch zur Echtzeitanalyse während des Spiels. Während es in anderen Sportarten bereits üblich ist, dass Trainer während des Spiels die gegnerische Mannschaft durch computergestützte Systeme analysieren (als Beispiel hierfür kann der American Football angeführt werden, wo Trainer aufbereitete Informationen über Tablet-PCs am Spielfeldrand erhalten), ist dies im Fußball noch unüblich. Jedoch gibt es hierzu bereits erste Testumgebungen

– so hat der Fußballverein TSG Hoffenheim in Zusammenarbeit mit dem Unternehmen SAP einen Teil seines Trainingszentrums diesbezüglich angepasst; hierbei befinden sich funkbasierte Sensoren im Ball und in den Trikots der Spieler, so dass sämtliche Bewegungen erfasst werden können (vgl. TSG 1899 Hoffenheim, 2012). In der Bundesliga ist der Einsatz derartiger Techniken jedoch (noch) verboten (vgl. Deutscher Fußball-Bund, 2015, S. 25).

Dennoch werden Spiele in Echtzeit analysiert, jedoch ohne die Zuhilfenahme von computergestützten Systemen, sondern durch die Anwendung der subjektiven Eindrucksanalyse durch die Trainerteams der Mannschaften, mit welcher folgende Rückschlüsse abgeleitet werden können (vgl. Bauer, 1998, S. 16):

- Anweisungen, die im Spielverlauf an Spieler weitergegeben werden können,
- Anweisungen in der Halbzeitbesprechung,
- Spielsystemanpassungen und Auswechslungen während des Spiels,
- Informationen für Spielbesprechungen nach dem Spiel,
- Konsequenzen für das Training und folgende Spiele.

In den folgenden Kapiteln erfolgt die Darstellung von Anwendungsgebieten, die nicht auf subjektiven Analysen basieren, sondern auf der systematischen Anwendung von Methoden auf Datenbeständen.

3.4.2 Leistungsbewertung von Spielern

Ein oftmals verwendetes Mittel zur Bewertung der Leistung von Spielern sind *Key Performance Indicators* (KPIs): diese Kennzahlen bestehen aus einer gewichteten Verrechnung verschiedener Merkmale, vergleichbar mit dem Ansatz von *Sabermetrics* im Baseball. Ein Beispiel für eine derartige KPIs ist der sog. Capello Index, der vom früheren englischen Nationaltrainer Fabio Capello im Rahmen seines Engagements bei der Weltmeisterschaft 2010 entwickelt wurde (vgl. Hamilton, 2010). Dieser basierte, Capello (2010a) zufolge, auf insgesamt etwa 100 verschiedenen Merkmalen, die verschiedene Events im Spiel abbilden (z. B. Pässe, Flanken, Schüsse, Zweikämpfe etc.); das Ergebnis ist ein Wert zwischen 0 und 100, wobei die besten Spieler der Weltmeisterschaft einen Wert von etwa 65 hatten (vgl. Capello, 2010b).

Ein weiteres Beispiel für den Einsatz der Leistungsbewertung ist der Trainer Arsene Wenger, der bereits in den späten 1980er Jahren beim französischen Verein AS Monaco auf computergestützte Analysen gesetzt hat, indem er das Programm *Top Score* eingesetzt hat (vgl. Kuper, 2011, S. 305). Diese Programm hat Kuper zufolge jedem Spieler Punkte für Aktionen innerhalb von Spielen zugewiesen, diese Punktestände hat der Trainer dann für seine Entscheidungen verwendet; Wengers Aussage zufolge hatten Spieler mit hohen Punktetständen später mit einer großen Wahrscheinlichkeit erfolgreiche Karrieren.

Mit diesen Kennzahlen kann nun eine Bewertung der Leistung von Spielern durchgeführt werden, wobei diese Kennzahlen sowohl für einzelne, als auch, wie oben dargestellt, für mehrere Spiele oder ganze Turniere berechnet werden können.

3.4.3 Identifikation von Erfolgsindikatoren

Neben der Verwendung von KPIs zur Bewertung von Spielern gab es im Fußball auch Versuche, erfolgsrelevante Eigenschaften im Bezug auf die Spielweise einer Mannschaft zu identifizieren. Beim bereits genannten Beispiel des dänischen Vereins FC Midtjylland wurde beispielsweise durch Analysen herausgefunden, dass Standardsituationen (d. h. Ecken, Freistöße und Einwürfe) die einfachste Möglichkeit sind, um Tore zu erzielen – daraufhin wurden Spieler verpflichtet, die im Rahmen solcher Standardsituationen überdurchschnittlich gute Leistungen erbringen konnten (vgl. Kjäll, 2015). Die Umsetzung dieser neuen Erkenntnisse verlief Kjäll zufolge sehr gut, inzwischen erzielt das Team von Midtjylland durchschnittlich ein Tor pro Spiel durch Standardsituationen, was europaweit ein Spitzenwert ist.

Ein weiteres Beispiel für die Erforschung von Erfolgsindikatoren ist die Arbeit von Lago, Lago und Rey (2011), die versucht haben, typische Eigenschaften erfolgreicher und erfolgloser Mannschaften herauszufinden. Das Ergebnis dieser Analyse war, dass *winning teams* deutlich bessere Werte hatten was die Anzahl von Schüssen, Torchancen-Verwertung und Passgenauigkeit anging, während *losing teams* tendenziell mehr Fouls und demnach mehr gelbe und rote Karten hatten (vgl. Lago u. a., 2011, S. 135). Eine konkrete Umsetzung der Erkenntnisse dieser Arbeit im Profifußball ist jedoch nicht bekannt, allerdings geben generell nur sehr wenige Vereine Informationen über derartige Erkenntnisse nach außen, da dies von konkurrierenden Vereinen verwendet werden könnte.

3.4.4 Kaderplanung

Ein weiteres Beispiel für die Verwendung von Datenanalysen ist die Kaderplanung, bei der folgende Aufgaben anfallen:

- Verpflichtung neuer Spieler,
- Auswahl von Spielern, die abgegeben werden sollen,
- Analyse des Entwicklungspotentials eigener (Jugend-)Spieler.

Da es sich hierbei um Beträge von mehreren Millionen Euro handelt (einmalige Zahlung an den abgebenden Verein und laufende Gehaltszahlungen an den Spieler), ist es logischerweise zu vermeiden, Spieler zu verpflichten, die sich im Nachhinein als Fehlentscheidung herausstellen, weshalb sich dieser Bereich sehr gut für Datenanalysen eignet.

Hierzu können KPIs einerseits dazu verwendet werden, um „schlechte“ Spieler in der eigenen Mannschaft zu identifizieren, die dann „verkauft“ werden können; andererseits können mit diesen auch überdurchschnittlich gute Spieler fremder Mannschaften identifiziert werden, die dann verpflichtet werden können. Beispiele hierfür lassen sich englischen Premier League finden, wo Anderson und Sally (2013, S.18) zufolge im Jahr 2011 die beiden Spieler Henderson und Downing aufgrund ihrer deutlich überdurchschnittlichen Zweikampfquoten im gegnerischen Strafraum vom Verein FC Liverpool verpflichtet worden seien, da diese Kennzahlen für Erfolg in der Pressing-Taktik stehen.

3.4.5 Medizinische Datenanalysen

Dann wird Data Mining heutzutage auch im Bereich der medizinischen Analysen eingesetzt, da die Verletzungen von Spielern möglichst minimiert werden sollten. Ein Beispiel hierfür ist der italienische Verein AC Mailand, wo Flinders (2002) zufolge im Jahr 2002 im Rahmen eines Prototyps versucht wurde, durch die Analyse physischer Daten sich anbahnende, unerkannte oder wiederkehrende Verletzungen herauszufinden und Wahrscheinlichkeiten für neue Verletzungen zu errechnen. Hierzu wurden Analysen auf Basis bestehender Datensätze der vorgehenden Jahren und bekannter Verletzungen der Spieler durchgeführt, die Genauigkeit der Erkennung von zukünftigen Verletzungen lag bei etwa 70 %.

3.4.6 Weitere Anwendungsmöglichkeiten

Darüber hinaus gibt es viele weitere Anwendungsmöglichkeiten, die zwar einen sportwissenschaftlichen Hintergrund besitzen, jedoch nur eine geringe Aussagekraft für das operative Geschäft im Fußball besitzen. So haben diverse Autoren Untersuchungen bezüglich der Klassifikation von Spielsystemen durchgeführt, beispielsweise hat Oberstone (2011) einen Vergleich der Spielweisen von Mannschaften der englischen, italienischen und spanischen Ligen durch die Anwendung von Regressionsanalysen vorgenommen.

Ein weiteres Beispiel sind Prognosen von Spielausgängen, die für die Mannschaften bzw. deren Trainer einen recht geringen Nutzen haben, da diese mehr an der Leistung der Mannschaft interessiert sind. Dennoch gibt es in diesem Bereich sehr viele Forschungsarbeiten, da Prognosen für Unternehmen außerhalb der Vereine sehr interessant sind, insbesondere für Wettanbieter, da deren Geschäftsmodell auf der Berechnung von Wettquoten basiert.

3.5 Probleme

Grundsätzlich ist es immer von der Sportart abhängig, ob sich der Einsatz von Data-Mining-Technologien in diesen eignet und ob es sich durchsetzt. Einige Sportarten sind diesbezüglich recht fortgeschritten, während sich andere noch am Anfang befinden. Schumaker u. a. (2010, S. 14) zufolge sind kommerzialisierte Sportarten, in denen es um große Geldbeträge geht, tendenziell offener für die Unterstützung von Entscheidungsprozessen durch Data Mining. Nichtsdestotrotz gibt es auch einige Kritikpunkte, die im Folgenden genannt werden, wobei sie sich auf den allgemeinen Einsatz im Sport beziehen; es werden jedoch auch Beispiele für den Fußball genannt.

Im Sport geht es nicht nur um quantitative Kennzahlen, da auch nicht-messbare Eigenschaften einen großen Einfluss haben – beispielsweise können hier psychologische Eigenschaften der Spieler genannt werden. Von Trainern und Spielern wird hier auf Eigenschaften wie Leidenschaft, Teamgeist und Fairness von Spielern verwiesen, die logischerweise nur schwer messbar und kaum beeinflussbar sind (vgl. Schumaker u. a., 2010, S. 14 f.).

Bei der Umsetzung müssen passende Kennzahlen ausgewählt werden – anfangs im Baseball gab es hier Probleme, da sich große Teile auf bewährte Kennzahlen fokussiert hatten, während sich der *Sabermetrics*-Ansatz an neuen, komplexeren Kennzahlen orientierte. Nun sind einfache Kennzahlen nicht zwingend falsch, jedoch sollte statt der „blinden" Verwendung historisch bewährter Kennzahlen der Zusammenhang zwischen Erfolg und diesen Kennzahlen analysiert werden (z. B. durch Regressionsanalysen). Beispielsweise werden defensive Spieler im Fußball oftmals anhand ihrer Zweikämpfe bewertet – nun hatte aber Paolo Maldini, ein erfolgreicher italienischer Nationalspieler, nur sehr selten Zweikämpfe geführt, weshalb er von seinem Trainer kritisiert wurde (vgl. Kuper und Szymanski, 2012, S. 166 f.). Jedoch war Maldinis Spielweise, wie Kuper und Szymanski ausführen, anders als die der anderen Defensivspieler, da er vorausschauender agierte und Zweikämpfe negativ einschätzte und diese deshalb vermied, da sie ein Zeichen dafür seien, dass Spieler aufgrund von Fehlern unter Druck gerieten.

Jedoch wird auch die Verwendung komplexerer Kennzahlen (KPIs) kritisch gesehen; grundsätzlich wäre das Ziel, dass menschliche Einflussfaktoren in Entscheidungsprozessen eliminiert werden (vgl. Stufe 5 des Stufenmodells der Verwendung von Daten im Sport, siehe Abschnitt 3.1.1). Jedoch sind die Ansätze von Data Mining im Sport nicht komplett objektiv, da die KPIs auf der gewerteten Verrechnung verschiedener Kennzahlen basieren. Hierfür muss die Gewichtung der verschiedenen Attribute festgelegt werden. Im Fall des Capello-Index (vgl. Abschnitt 3.4.2) wurde dies vom Trainer Capello in Zusammenarbeit mit Experten gemacht; grundsätzlich wäre es auch möglich, derartige Indizes komplett objektiv zu erstellen, indem die Gewichtung automatisiert durchgeführt wird, was aber mit einem deutlich größeren Aufwand verbunden wäre (vgl. Hamilton, 2010).

Ein wichtiger Punkt ist zudem die Generalisierbarkeit von Anwendungsszenarien, da im Fußball eine große Abhängigkeit zu den konkreten Teams vorhanden ist, so dass Kennzahlen, die in bestimmten Vereinen als aussagekräftig bestimmt wurden, in anderen Vereinen nur eine geringe Aussagekraft besitzen. Das folgende Zitat von Medeiros (2014) beschreibt dieses Problem anhand des englischen Vereins *Manchester City*:

> Statistics such as line breaks and possession in the last third are important for [Manchester] City but would probably be irrelevant to a team with a different style: football analytics is a discipline in which the way a team plays dictates which statistics are significant. The challenge is to find out which.

Die Darstellung der Einsatzmöglichkeiten von Data Mining im Sport – und insbesondere im Fußball – ist hiermit abgeschlossen; im nächsten Kapitel erfolgt die Betrachtung der im Hintergrund von Data Mining verwendeten Datengrundlage.

4 Datengrundlage

Wie bereits im vorherigen Kapitel dargelegt wurde, besteht die größte Herausforderung beim Einsatz von Datenanalysen im Fußball darin, geeignete Kennzahlen zu finden, die möglichst aussagekräftig sind und einen Zusammenhang mit dem Erfolg von Mannschaften haben sollten. Es gibt jedoch eine Vielzahl an Kennzahlen, außerdem können die Datenmengen sehr groß werden, wenn Analysen sich auf allgemeine Kennzahlen fokussieren. So können pro Fußballspiel mehr als 60 Millionen Datensätze erzeugt werden, wenn sämtliche Positionsdaten und Geschwindigkeiten der einzelnen Spieler erfasst werden; alternativ können Analysen jedoch auch aggregierte Kennzahlen verwenden, wie beispielsweise Ballbesitzwerte oder Passgenauigkeiten (vgl. FC Bayern München AG, 2015).

Neben der Darstellung und Bewertung verschiedener Kennzahlen erfolgt in diesem Kapitel auch die Differenzierung möglicher Analyseziele, zudem wird auf Technologien zur Erfassung dieser Kennzahlen und auf mögliche Datenquellen eingegangen.

4.1 Differenzierungen

Grundsätzlich muss zuerst darüber nachgedacht werden, was das Ergebnis der Analysen sein soll, da die verschiedenen Möglichkeiten unterschiedliche Input-Kennzahlen zur Durchführung benötigen. Data Mining kann an dieser Stelle eine von vielen Möglichkeiten zur Durchführung sein; jedoch können Analysen beispielsweise auch durch Expertenmeinungen oder durch einfache statistische Aufbereitungen stattfinden. Im Folgenden wird eine Gegenüberstellung möglicher Varianten von Analysen dargestellt, beruhend auf den Überlegungen von Leser (2007, S. 14 ff.).

Qualitative vs. quantitative Analysen:
Grundsätzlich lassen sich Analysen anhand von qualitativen und quantitativen Merkmalen durchführen. Bei der Anwendung quantitativer Methoden werden Bauer (1998, S. 13f.) zufolge technische Aktionen gezählt, beispielsweise Torschüsse, Flanken und Pässe, die gleichzeitig in die Kategorien „erfolgreich“ und „nicht erfolgreich“ unterteilt werden. Um möglichst objektive Datensätze zu generieren, müssen nach Winkler (2000, S. 68) Aktionen im Vorhinein definiert werden, beispielsweise muss geklärt werden, wann eine Spielsituation als ein Zweikampf zwischen zwei Spielern angesehen werden kann und wie der Erfolg eines Zweikampfes definiert wird.

Da Bauer (1998, S. 14) zufolge hierbei auch aufgezeichnet wird, welche Spieler an den Aktionen beteiligt waren, an welcher Position auf dem Spielfeld und zu welchem Zeitpunkt die Aktionen stattfanden, kann hierdurch ein objektives Spielprotokoll erzeugt werden.

Im Gegensatz dazu versteht man unter einer qualitativen Spielanalyse gemäß Winkler (2000, S. 69) „sowohl die Analyse von Spielszenen hinsichtlich der Güte von Handlungen, des Verhaltens und Könnens einzelner Spieler, als auch die Analyse der Spielstruktur einer Mannschaft“. Während bei der quantitativen Analyse also nur die Ergebnisse der

einzelnen Aktionen (bspw. Torschuss, Pass, Zweikampf) betrachtet werden, kann hierbei auch die Qualität der Durchführung betrachtet werden; außerdem können Bauer (1998, S. 14) zufolge auch „komplexe gruppen- und mannschaftstaktische Handlungen“ erfasst werden.

Im Bezug auf die Datenerfassung innerhalb von Spielen wird in der Literatur oftmals auf die beiden Prozesse der systematischen Spielbeobachtung und der subjektiven Eindrucksanalyse verwiesen. Während nach Lames (vgl. 1994, S. 23 f.) die subjektive Eindrucksanalyse davon ausgeht, dass bereits während der Beobachtung eine „analytische Absicht“ vorhanden ist, da der analysierende Experte sein Fachwissen einbringt, werden bei der systematischen Spielbeobachtung während des Spiels nur die quantitativen Merkmale erfasst, die dann im Nachhinein (automatisiert) analysiert werden können (vgl. Tabelle 2).

Subjektive Eindrucksanalyse	**Systematische Spielbeobachtung**
flexible Merkmale	genau festgelegte Merkmale
ohne systematische Fixierung	systematische Fixierung
Eindrücke	Beobachtungen

Tabelle 2: Formen der Spielbeobachtung im Fußball nach Lames (1994, S. 24)

Die beiden verschiedenen Formen der Spielbeobachtung sind Lames (1994, S. 24) zufolge jedoch nicht als konkurrierende Ansätze zu verstehen, sondern als sich gegenseitig ergänzende Mittel zur Spielbeobachtung, da beide Ansätze verschiedene Vor- und Nachteile haben. So lassen sich bei der subjektiven Eindrucksanalyse auch sonst nicht-erfassbare Eigenschaften erfassen, als Beispiel nennt Lames hier die „Qualität von taktischen Handlungsentscheidungen“, während durch eine systematische Erfassung lediglich „Ausgangssituation und Resultat“ erfassbar wären.

Des Weiteren kann die Kombination der beiden Ansätze auch zur Aufdeckung von Schwächen verwendet werden; so können quantitative Analysen zur Identifizierung von Schwächen genutzt werden (z. B. Spieler X hat schlechte Zweikampfwerte), während qualitative Analysen einen „Aufschluss auf die Ursachen von Fehlern und Mängeln“ geben können (vgl. Winkler, 2000, S. 69).

Ereigniseinheiten vs. Zeiteinheiten:
Eine weitere Unterteilung ist hinsichtlich der Wahl der Beobachtungseinheiten notwendig, Lames (1994, S. 51 f.) zufolge kommt es im Bezug auf die Spielbeobachtung nur in Frage, komplette Spiele zu beobachten, jedoch muss der Spielverlauf von 90 Minuten auf erfassbare Einzelaktionen eingeschränkt werden. Hierzu gibt es nach Lames einerseits den Ansatz der Zeiteinheitentechnik, bei welcher Zeitintervalle definiert werden (z. B. 5 Sekunden), bei welchen der aktuelle Spielzustand erfasst wird; andererseits gibt es die Ereigniseinheitentechnik, im Rahmen derer von der Strukturierung des Spielverlaufs als Kette von Ereignissen ausgegangen wird, so dass nur die einzelnen Ereignisse erfasst werden. Im

Vergleich der beiden Techniken zeigt sich, dass sich mit der Zeiteinheitentechnik mehrere parallel laufende Aktionen erfassen lassen, deren „Zustandsänderungen von Zeitpunkt zu Zeitpunkt“ erfasst werden; im Gegensatz hierzu findet bei der Ereigniseinheitentechnik eine Beschränkung auf einzelne Handlungsstränge statt, da beispielsweise nur Ballaktionen erfasst werden (vgl. Leser, 2007, S. 15).

Allgemeingültige vs. individuelle Ergebnisse:
Eine weitere zu beantwortende Fragestellung ist angesichts der Ergebnisse der Analysen zu treffen:

- Sollen die Analysen allgemeingültige Ergebnisse liefern, indem die Daten mehrerer Ligen als Grundlage verwendet werden, z. B. zur Analyse der strukturellen Unterschiede der Spielsysteme der englischen und deutschen Liga, oder
- sollen nur Vereine innerhalb einer Liga analysiert werden, wie z. B. zur Bestimmung der Erfolgsfaktoren in der deutschen Bundesliga, oder
- sollen die Analysen nur eine einzelne Mannschaft im Fokus haben, z. B. zur Ermittlung der Wichtigkeit des Spielers X innerhalb einer Mannschaft?

Darüber hinaus ist auch die Analyse ohne einen konkreten Bezug auf jegliche Vereine oder Ligen möglich, dies sind jedoch oftmals grundsätzliche Forschungsgebiete, bei denen es beispielsweise um die Anwendung neuer Techniken im Fußball geht, z. B. die automatische Datenerfassung durch die Analyse von Videodaten.

Im Allgemeinen ist es fraglich, ob es überhaupt möglich ist, allgemeingültige Ergebnisse zu generieren, da die verschiedenen Ligen recht unterschiedlich sind; so sind einerseits die Spielsysteme recht unterschiedlich (z. B. ist die englische Liga recht offensiv und konterlastig, während die deutsche Liga tendenziell auf einer soliden Defensive mit geordnetem Spielaufbau basiert), anderseits sind recht große Qualitätsschwankungen möglich, sofern abseits der „großen“ Ligen in Europa analysiert werden soll.

Spieleranalyse vs. Mannschaftsanalyse:
Im Bezug auf die zu beobachtenden Merkmale lässt sich ebenfalls eine Differenzierung durchführen, da einerseits gesamte Mannschaften, andererseits aber auch nur einzelne Spieler betrachtet werden können. Beruhend auf den Ausführungen von Lange (1997, S. 36) lässt sich hier folgende Dreiteilung durchführen:

- **Individualtaktik**: Die Analyse kann grundsätzlich auf der feinsten Granularität auf Basis eines einzelnen Spielers stattfinden, beispielsweise anhand von Merkmalen wie Ballkontrolle, Laufwegen, Torschüssen und Pässen eines Spielers.
- **Gruppentaktik**: Zusätzlich zu den Daten eines einzelnen Spielers sollte jedoch auch das Zusammenspiel mit seinen Mitspielern betrachtet werden; im Bezug auf die Gruppentaktik können beispielsweise die Spieler des Angriffs, des Mittelfelds und der Abwehr getrennt betrachtet werden. Als Beispiel hierfür nennt Lange im Rahmen

des Angriffspiels das Kombinationsspiel zwischen Spielern oder das Umschaltspiel von Defensive auf Offensive.

- **Mannschaftstaktik**: Im Hinblick auf die Analyse der gesamten Mannschaft können sämtliche Mitglieder einer Mannschaft analysiert werden, Lange führt hier u. a. Standardsituationen oder das Verhalten von Spielern bei gegnerischem Ballbesitz an.

In der Literatur wird die Beobachtung einzelner Spieler kritisiert, da Fußball ein Mannschaftssport ist, weshalb die Leistung eines Spielers stark von seinen Mitspielern abhängig ist. Lames (1994, S. 27) beschreibt dies passend in folgendem Zitat:

> Die individuelle Leistung muss aus vorsichtigen Interpretationen des Mannschaftsverhaltens erschlossen werden. Die Ganzheitlichkeit des Interaktionsprozesses Sportspiel verbietet die völlig isolierte Betrachtung eines Spielers. Ohne gleichzeitige Berücksichtigung der Mitspieler und des Gegners erhält man kein korrektes Abbild der Sportspielrealität.

Als Beispiel hierfür können Stürmer genannt werden, deren Erfolge stark von ihren Mitspielern und auch von den gegnerischen Defensivspielern und dem gegnerischen Torwart abhängig sind, da sie einerseits auf Pässe von Mitspielern angewiesen sind, andererseits kann es z. B. auch sein, dass der Torwart sämtliche Torschüsse hält, obwohl die Leistung des Stürmers vergleichsweise gut ist.

Prognosen vs. Performance Analysis:
Grundsätzlich gibt es zwei verschiedene Einsatzmöglichkeiten von Datenanalysen im Fußball: einerseits können Algorithmen zur Prognose eingesetzt werden, andererseits kann das „klassische“ Data Mining auf Beschreibungsprobleme bzw. zur Mustererkennung angewandt werden. Im Rahmen von Prognosen wird primär versucht, Ergebnisse von einzelnen Spielen oder von ganzen Spielzeiten vorherzusagen, während sich bei der Mustererkennung deutlich komplexere Einsatzmöglichkeiten ergeben, da dies im Rahmen der *Performance Analysis* basierend auf historischen Daten verwendet werden kann. Direkt innerhalb von Fußballvereinen wird Data Mining daher vor allem zur Analyse der Leistung verwendet, mit dem Ziel des Schaffens von Wettbewerbsvorteilen gegenüber gegnerischen Mannschaften. Dies kann einerseits durch die Analyse der eigenen Leistung stattfinden (und durch die Eliminierung eigener Schwächen); andererseits durch die Analyse der gegnerischen Mannschaften und ihrer Stärken und Schwächen.

4.2 Kennzahlen

Grundsätzlich gibt es im Fußball viele verschiedenen Kennzahlen, die ihren Ursprung entweder in den Spielen oder aber auch in der Spielvorbereitung bzw. im Training haben können. So ist es bei professionellen Fußballvereinen heutzutage üblich, dass im Training sowohl Kamerasysteme zur Erfassung physischer Daten (z. B. Laufdaten oder Pässe), als

auch Sensoren zur Erfassung medizinischer Daten (z. B. Puls oder Atemfrequenz) verwendet werden (vgl. Grüling, 2015). Im Folgenden erfolgt jedoch eine Konzentration auf die Datenerfassung innerhalb von Spielen, da nur recht wenige Details über die Erfassung innerhalb des Trainings nach außen gelangen.

4.2.1 Spielergebnisse

Grundsätzlich ist es naheliegend, die Ergebnisse der einzelnen Fußballspiele als Datengrundlage zu verwenden, da diese die Grundlage der jeweiligen Fußball-Ligen darstellen und frei verfügbar sind. Hier gibt es einerseits Kennzahlen auf Basis einzelner Spiele (Spielergebnisse, Tore, Gegentore), andererseits kann auch die Ergebnistabelle einer Liga verwendet werden. Im Fußball wird der Erfolg von Mannschaften anhand ihrer Tabellenposition beurteilt – die Tabelle ist hierbei eine aggregierte Darstellung der einzelnen Spiele einer Saison (vgl. Tabelle 3). Neben der Anzahl der jeweiligen Spielergebnisse (Siege, Unentschieden, Niederlage) und der Punktzahl wird auch die Anzahl aller erzielten Tore und erhaltenen Gegentore sowie die Tordifferenz abgebildet.

Platz	**Club**	**Spiele**	**S**	**U**	**N**	**Tore**	**Tordifferenz**	**Punkte**
1	FC Bayern München	34	28	4	2	80:17	+ 63	88
2	Borussia Dortmund	34	24	6	4	82:34	+ 48	78
3	Bayer Leverkusen	34	18	6	10	56:40	+ 16	60
...								
16	Eintracht Frankfurt	34	9	9	16	34:52	– 18	36
17	VfB Stuttgart	34	9	6	19	50:75	– 25	33
18	Hannover 96	34	7	4	23	31:62	– 31	25

Tabelle 3: Ausschnitt Ergebnistabelle Bundesliga Saison 15/16 (vgl. Deutsche Fußball Liga GmbH, 2015)

Neben der abgebildeten einfachen Tabelle gibt es weitere Formen, beispielsweise werden die verschiedenen Kennzahlen im Rahmen der Heim- und Auswärtstabellen nach Heim- bzw. Auswärtsspiel untergliedert; eine weitere Variante ist die Formtabelle, in welcher nur die letzten fünf Spiele betrachtet werden, um die aktuelle Form der jeweiligen Teams beurteilen zu können.

Neben dieser aggregierten Form können die Spielergebnisse auch einzeln betrachtet werden, wobei sich hier einerseits die Leistung der Mannschaften und einzelner Spieler betrachten lässt (vgl. Abschnitt 4.2.2), andererseits lassen sich auch grundsätzliche Details zu den Spielen erfassen, was in der folgenden Aufzählung dargestellt wird (vgl. ZDF Sport, 2015):

- beteiligte Mannschaften,
- Trainer der Mannschaften,

- Datum und Uhrzeit,
- Spielort und Name des Stadions,
- Anzahl der Zuschauer,
- Schiedsrichter,
- Aufstellung,
- Spielergebnis,
- Spielverlauf (Tore, gelbe und rote Karten, Ein- und Auswechslungen).

4.2.2 Spieldaten

Neben den genannten grundsätzlichen Daten zu Spielen lassen sich aber auch deutlich detailliertere Informationen erfassen. Hier lässt sich eine Unterteilung in zwei verschiedene Kategorien vornehmen, einerseits in die *physischen* Daten, die einen Bezug zur Position und zur Bewegung von Spielern unabhängig vom Ballbesitz haben, anderseits in *taktische* Daten, was mit Ballereignissen, d. h. Aktionen von Spielern mit dem Ball gleichzusetzen ist (vgl. Biermann, 2011, S. 60 f.). Im Folgenden werden die beiden Kategorien kurz erläutert und deren verschiedene Kennzahlen angegeben.

Physische Daten
Unter dem Aspekt der physischen Daten werden Ereignisse betrachtet, die keinen direkten Bezug zum Ball im Spiel haben. Dies sind hauptsächlich Positionsdaten von Spielern, d. h. welcher Spieler befindet sich gerade an welchem Ort auf dem Spielfeld. Hier lassen sich, je nach Wahl des Zeitintervalls, sehr viele Datenpunkte erzeugen. Würde man beispielsweise für jeden Spieler im Abstand von drei Sekunden die Position erfassen, könnten im Laufe eines Spiels (90 Minuten, 11 Spieler) etwa 20.000 Datensätzen erfasst werden.

Taktische Daten
Hierunter fallen nun, wie bereits dargestellt, sämtliche Aktionen, die einen Bezug zum Ball besitzen, daher gibt es in dieser Kategorie im Vergleich zu den physischen Daten, wo grundsätzlich nur Positionen von Spielern erfasst werden, eine deutlich größere Menge an verschiedenen Kennzahlen. Die Kennzahlen des kommerziellen Statistikanbieters Opta Sports (vgl. Kapitel 4.3) lassen sich auf folgende Weise aufgliedern, wobei sich diese Aufzählung auf zentrale Elemente beschränkt, da Opta Sports insgesamt über 60 verschiedene Kennzahlen bereitstellt (vgl. Opta Sports, 2013b):

- Ballkontakte,
- Pässe,
- Schüsse,
- Tore, Torvorlagen,
- Standardsituationen (Ecken, Freistöße, Elfmeter),
- Torwartparaden,
- Zweikämpfe, Fouls, gelbe Karten, rote Karten.

Diese verschiedenen Bereiche unterteilt Opta Sports in viele einzelne Kennzahlen, wie an der folgenden beispielhaften Darstellung der Kategorie Pässe zu sehen ist. Hierbei werden Opta Sports (2013b) zufolge bei jedem Pass die x- und y-Koordinaten des Passgebers und des Empfängers erfasst, wodurch Analysen im Bezug auf die Länge des Passes, im Bezug auf die Richtung des Passes und im Bezug auf die Spielfeldunterteilung ermöglicht werden. Grundsätzlich gibt es also bei Pässen folgende Unterteilung:

- erfolgreiche oder nicht erfolgreiche Pässe,
- lange oder kurze Pässe,
- Pässe in der eigenen oder gegnerischen Spielfeldhälfte,
- Pässe im eigenen oder gegnerischen Strafraum,
- gelupfte Pässe,
- Flanken,
- Kopfballpässe,
- Rückpässe,
- uvm.

Im Fußball lässt somit sich eine sehr große Menge an Datensätzen erzeugen, Biermann (2011, S. 53) zufolge werden in einem Spiel rund 2500–3000 „ballgebundene“ Aktionen erfasst, außerdem können, wie schon dargestellt, ebenfalls mehrere Tausend „laufgebundene“ Aktionen erfasst werden. Eine derartig hohe Anzahl an Datensätzen ist natürlich für das menschliche Auge recht unübersichtlich, außerdem können Analysen, die auf mehreren Spielen oder ganzen Spielzeiten basieren, recht komplex werden. Daher erzeugt man aus diesen Daten aggregierte Kennzahlen, beispielsweise werden aus den physischen Daten Bewegungsprofile, durchschnittliche Geschwindigkeiten oder Laufdistanzen errechnet, aus den taktischen Daten können beispielsweise die Anzahl von Torschüssen, Ballbesitzwerte oder Passgenauigkeiten errechnet werden (ein Beispiel hierfür siehe ZDF Sport, 2015).

4.2.3 Qualitative Kennzahlen

Neben den beiden genannten Kategorien, die quantitativ sind, gibt es auch den Bereich der qualitativen Kennzahlen. Nopp (2012, S. 97) zufolge ist hierbei die „trennscharfe Kategorisierung von Variablen [...] ungleich schwieriger als bei der Definition quantitativer Kriterien“. Es werden grundsätzlich die zwei Bereiche des Angriffs- und des Abwehrverhaltens unterschieden, wobei zum ersteren beispielsweise Werte wie Spielgeschwindigkeit, Spielkontrolle, Risiko und Spielverlagerung angehören, während letztere aus den Teilen Anlaufen und ballorientiertes Stören besteht (vgl. Nopp, 2012, S. 97 f.). Eine weitere Möglichkeit des Einsatzes von qualitativen Kennzahlen ist die Verwendung der im Abschnitt 3.4.2 genannten KPIs (z. B. Capello-Index), welche aus einer Kombination mehrerer Merkmale bestehen und zur Leistungsbewertung verwendet werden können.

Eine detaillierte Darstellung dieses Bereiches erfolgt an dieser Stelle nicht, da die Erfassung derartiger Kennzahlen mit einem größeren Aufwand und auch der vorherigen

Erarbeitung ihrer sportwissenschaftlichen Definitionen verbunden ist, weshalb sich diese Arbeit grundsätzlich auf die quantitative Betrachtung des Fußballes konzentriert.

4.3 Datenquellen

Als Datenquellen werden im Fußball – neben den Daten, die von den Fußballvereinen selbst erfasst werden (z. B. im Training) – meist *kommerzielle Anbieter* verwendet, hierbei sind die größten Anbieter die drei Firmen Opta Sports, Impire und Prozone. Die folgende Darstellung beschränkt sich auf den Anbieter Opta Sports, da diese als einziges Unternehmen recht ausführliche Beschreibungen öffentlich zugänglich machen; jedoch sind sich die verfügbaren Daten der drei Anbieter im Grundsatz recht ähnlich.

Opta Sports ist eigenen Aussagen zufolge der führende Anbieter von Sportdaten und bietet Daten für verschiedene Sportarten an, unter anderem Fußball, Rugby, Kricket, Tennis und American Football (vgl. Opta Sports, 2016c). Im Bereich Fußball hat Opta Sports viele namhafte Kunden, die sich den Kategorien Medien, Fußballvereine und -Verbände und Sponsoren zuordnen lassen, Beispiele hierfür sind z. B. ZDF, Sky, Bild, Kicker, Adidas, Nike, sowie zahlreiche Fußballvereine und Verbände der großen europäischen Ligen wie z. B. 12 der 18 Bundesligisten oder der DFB (vgl. Opta Sports, 2016b).

Außerdem ist Opta Sports seit der Saison 2013/2014 offizieller Datenlieferant der deutschen Bundesliga und bietet hierbei folgende Dienste an (vgl. Opta Sports, 2013a):

- XML-Datenfeeds,
- Redaktionelle Leistungen (z. B. Spielfakten und Spielerprofile),
- Widgets (Einbindung aufbereiteter Daten in Webseiten),
- Match Center (Live-Ticker),
- Grafiken für Print- und Online-Medien.

Hieraus wird ersichtlich, dass die verschiedenen kommerziellen Anbieter nicht nur einfache Daten liefern, sondern auch aufbereitete Statistiken, die z. B. von den Medien direkt verwendet werden können. Im Bezug auf die Datenfeeds, die im Rahmen des Data Mining von größtem Interesse sind, bietet Opta Sports folgendes an (vgl. Opta Sports, 2016c):

- **Core Data Feeds:** grundsätzliche Informationen über Spiele, wie z. B. Aufstellungen und Ergebnisse,
- **Classic Data Feeds:** detailliertere Statistiken, die sich sowohl auf einzelne Spieler, als auch auf gesamte Mannschaften beziehen können, wie sie bereits im Abschnitt 4.2.2 in der Kategorie der taktischen Daten dargestellt wurden,
- **Performance Data Feeds:** sehr detaillierte Daten, einschließlich der x- und y-Koordinaten und Zeitpunkte der einzelnen Aktionen.

Welche Preise Opta Sports für diese Dienste aufruft, wird von diesen nicht öffentlich gemacht, jedoch finden sich im Internet verschiedene Zahlen von Kunden: so soll Opta

Sports je nach Art und Umfang des Datenfeeds Preise beginnend von 1.000 $ bis hin zu 11.000 $ pro Saison verlangen (vgl. Reisenweber, 2016).

Neben den kommerziellen Anbietern gibt es heutzutage zahlreiche *frei verfügbare Quellen* im Internet, wozu einerseits Fußballwebseiten zählen, die im Rahmen ihrer Berichterstattung Sportstatistiken darstellen, und andererseits freie Datenbanken, die Fußballergebnisse und -statistiken in textbasierten Dateiformaten bereitstellen.

Da die Fußballwebseiten oftmals recht viele Leser haben, aktualisieren diese ihre Statistiken recht zuverlässig nach den jeweiligen Spieltagen, außerdem greifen diese größtenteils auf die oben genannten kommerziellen Anbieter zurück, was aus den Auflistungen der Kunden der Statistikanbieter hervorgeht. Als Beispiel für derartige Seiten können `transfermarkt.de`, `spox.de` oder `bundesliga.de` genannt werden.

Im Gegensatz hierzu werden die freien Datenbanken oftmals als Hobby-Projekte von Statistikern durchgeführt, so dass diese teilweise unzuverlässig aktualisiert werden und vor allem nur eine recht geringe Anzahl an Kennzahlen bereitstellen. Als Beispiel hierfür kann das Projekt `openfootball.github.io` genannt werden, das Daten von vielen europäischen Ligen und Wettbewerben bereitstellt, beispielsweise die englische, deutsche, italienische und spanische Liga, außerdem die Europa- und Weltmeisterschaften und viele weitere Wettbewerbe. Im Bezug auf die Menge an betrachteten Spielzeiten ist dieses Projekt sehr groß, da nicht nur aktuelle, sondern auch historische Daten vorhanden sind, jedoch sind für die einzelnen Spiele nur die beteiligten Mannschaften und die Spielergebnisse vorhanden, nicht aber detailliertere Leistungsdaten der Mannschaften.

Während die Daten durch die kommerziellen Anbieter oftmals in XML-Form bereitgestellt werden, verwenden die freien Datenbanken oftmals einfache Textform (d. h. keine standardisierte Form) oder die Formate CSV oder JSON. Letztere sind für die maschinelle Weiterverarbeitung durch das Data Mining geeignet, was bei den Fußballwebseiten nicht ohne weiteres möglich ist, da diese Statistiken in die HTML-Webseiten eingebettet sind. Hierzu können jedoch Crawler (*web scraping*) verwendet werden, um diese zu extrahieren, was jedoch mit Programmieraufwand verbunden ist.

Je nach Auswahl des Einsatzbereiches beim Data-Mining-Einsatz muss daher abgewogen werden, ob ausführliche Daten benötigt werden, oder ob Spieldaten und aggregierte Kennzahlen auf Teambasis genügen. Betrachtet man die im Kapitel 3.4 dargestellten Einsatzmöglichkeiten, würde die Leistungsbewertung einzelner Spieler ohne detaillierte Spieldaten auf Basis einzelner Spieler nicht möglich sein, während beispielsweise eine Prognose von Ergebnissen auch auf den reinen Spielergebnissen und grundsätzlichen Spieldaten möglich ist.

4.4 Datenerfassung

Die einfachste Möglichkeit, Daten von Fußballspielen ohne technische Hilfsmittel zu erfassen, ist die **manuelle Erfassung**, wie sie früher üblich war. So empfahl beispiels-

weise Bauer (1998, S. 14) die schriftliche Erfassung anhand von Vordrucken und Live-Tonaufzeichnunung mit Tonbändern. Heutzutage wird dieser Ansatz noch immer verwendet, jedoch werden hierbei mehr technische Hilfsmittel verwendet – so wird bei Opta Sports ein Programm verwendet, um die Events möglichst schnell in Echtzeit registrieren zu können, wobei drei Mitarbeiter pro Spiel eingesetzt werden (vgl. Opta Sports, 2016a):

> Drei Experten sind in jedem Spiel aktiv, wobei die Aktionen der Heimmannschaft von einem und die des Gästeteams von einem anderen Analysten aufgezeichnet werden, während ein dritter die erfassten Daten auf Konsistenz überprüft und mit weiteren Informationen ergänzt.

Heutzutage gibt es jedoch auch fortgeschrittenere Möglichkeiten als die manuelle Erfassung. So können Ansätze der künstlichen Intelligenz zur **automatischen Bilderkennung** verwendet werden, was verschiedene kommerzielle Anbieter in der den „Top-Ligen“ in Deutschland, England, Spanien und Italien umsetzen. Beim Anbieter Prozone werden bei einer Installation deren Systems acht Kameras in den Stadien eingebaut, wobei die Kameras so positioniert werden, dass jeder Punkt auf dem Spielfeld von mindestens zwei Kameras erfasst wird, so dass eine zweidimensionale Abbildung des Spiels durchgeführt werden kann (vgl. Di Salvo u. a., 2006, S. 111).

Biermann (2011, S. 55) zufolge sind derartige Systeme jedoch nicht zu 100 % exakt, da hierbei grundsätzlich nur die Positionen und Bewegungen der Spieler registriert werden, so dass die Erkennung komplexerer Events durch Berechnungen erfolgen muss. Als Beispiel hierfür nennt Biermann Zweikämpfe zwischen Spielern um den Ball, da das System hierbei erkennen muss, ob ein Kontakt ein Zweikampf ist und welcher Spieler der Gewinner des Zweikampfes ist; hier sei daher eine manuelle Aufbereitung durch Analysten notwendig.

Eine weitere Möglichkeit zur Datenerfassung ist der Einsatz von **Sensoren**, was beispielsweise vom Fraunhofer Institut durch das System *RedFIR* umgesetzt wurde, bei welchem sich Funksender im Ball und an den Spielern zur Positionserfassung befinden (vgl. Fraunhofer Institut für Integrierte Schaltungen, 2016). Während üblicherweise nur ein Datenpunkt pro Sekunde erfasst wird, kann hier eine deutlich größere Datenmenge erzeugt werden. So wurden bei *RedFIR* die Positionen der Spieler 200 mal pro Sekunde (je 200 Datensätze pro Fuß, bei den Torwärten außerdem Erfassung der Handpositionen) und die Position des Balles 2000 mal pro Sekunde erfasst, für ein ganzes Spiel lassen sich daher über 60 Mio. Datenpunkte im dreidimensionalen Raum erfassen (vgl. Sumpter, 2016, S. 167 f.). Da das System nur zur Erkennung von sog. „Basisevents“ und zur Erfassung physischer Daten vorgesehen ist, sind die Probleme von Systemen zur Bilderkennung (z. B. Erkennung von Zweikämpfen) hierbei auch vorhanden (vgl. Fraunhofer Institut für Integrierte Schaltungen, 2016).

5 Anwendung von Data Mining zur Prognose

Für den Einsatz von Data Mining gibt es, wie im Abschnitt 3.4 dargestellt, auch im Sport einige Einsatzmöglichkeiten, wobei sich viele allerdings im Bereich der *Performance Analysis* befinden und somit einen recht großen sportwissenschaftlichen Anteil besitzen. Da die vorliegende Arbeit thematisch der Wirtschaftsinformatik zuzuordnen ist, würden sich diese Bereiche eher nicht zur Bearbeitung eignen. Daher wurde die Anwendung von Data Mining zur Prognose von Spielergebnissen ausgewählt, da hier nur geringe sportwissenschaftliche Kenntnisse von Nöten sind; außerdem eignet sich die Prognose von Spielen, Heuer und Rubner (2013, S. 8) zufolge, für die Anwendung von Analysen, da wohldefinierte Daten frei verfügbar sind und das Interesse in diesem Bereich recht groß ist. Die Spielergebnisse sind hierbei nicht nur für die einzelnen Vereine interessant, sondern unter anderem auch für die Fans, die mit ihren Mannschaften „mitfiebern".

Statistische Modelle von Prognosen werden unter der Annahme entwickelt, dass die betrachteten Spiele zu „korrekten" Spielergebnissen führen, d. h. dass die „bessere" Mannschaft gewinnt, wobei dies unter anderem anhand der Qualität der Mannschaft oder anhand der Spielleistung festgemacht werden kann (vgl. Skinner und Freeman, 2009, S. 1087). Hierbei liegt ein Akzent dieser Arbeit darauf, die Qualität einer Mannschaft durch Zahlen zu definieren, was aufgrund der großen Menge an verfügbaren und errechenbaren Kennzahlen recht komplex ist; im Abschnitt 5.5 wird dies durch die Anwendung einer Korrelationsanalyse zu lösen versucht.

Ein weiteres Problemfeld ist, dass der Zufall einen recht großen Einfluss auf die Spielergebnisse hat – dem Zufall lassen sich hierbei Einflussfaktoren zuschreiben, die – anders als bspw. die „Qualität der Mannschaften, aktuelle Form, mentale Stärke, Verletzungen oder Sperren" (vgl. Quitzau und Völpel, 2009, S. 15) – nicht prognostizierbar sind. Im Falle eines Schusses kann sich der Zufall nach Lames (1999, S. 144 f.) beispielsweise in Pfosten- und Lattenschüssen, Torwartberührungen oder abgefälschten Schüssen manifestieren. Laut einer Analyse von Anderson und Sally (2013, S. 53) gewinnt im Fußball nur in etwa 50 % der Spiele die Mannschaft, die im Vorfeld als Favorit ermittelt wurde; dieser Wert ist anderen Sportarten deutlich höher, im Handball, Baseball und American Football gewinnt in etwa 70 % der Spiele der Favorit. Dies lässt erkennen, dass der Zufall eine recht große Bedeutung im Fußball hat. Ein bekanntes Beispiel dafür, dass nicht immer die „bessere" Mannschaft gewinnt, ist das Champions-League-Finale 2012, in welchem der FC Bayern München auf den englischen Verein FC Chelsea aus London traf. Hierbei hatte der FC Bayern München in nahezu allen Kennzahlen deutlich bessere Werte, da sich ihr Gegner auf die Defensive fokussierte; beispielsweise standen den 35 Torschüssen und 20 Eckbällen von Bayern nur 9 Torschüsse und ein Eckball von Chelsea gegenüber; trotzdem schaffte es Chelsea, durch ein Tor nach ihrem einzigen Eckball glücklich den Ausgleich und damit das Elfmeterschießen zu erreichen, welches sie anschließend gewannen (vgl. Anderson und Sally, 2013, S. 39).

Im Rahmen dieser Arbeit sollen Modelle erzeugt werden, die eine möglichst genaue Prognose ermöglichen. Aufgrund des bereits genannten Einflusses des Zufalls wird diese Genauigkeit wohl recht weit von den maximal erreichbaren 100 % entfernt sein. Grundsätzlich liegen derartig erzeugte Modelle im Bezug auf ihre Genauigkeit über den Vorhersagen von Laien und Experten, sie haben jedoch eine geringere Aussagekraft als die der Buchmacher von Sportwetten, wobei letzteres darin begründet ist, dass die Modelle der Buchmacher einerseits aussagekräftigere Informationen als Input verwenden und dass sie andererseits deutlich mehr Aufwand in ihren Analyseprozessen aufwenden können, so dass diese deutlich komplexer und effizienter sind (vgl. Heuer und Rubner, 2012, S. 1).

Der Aufbau des vorliegenden Kapitels orientiert sich grundsätzlich am KDD-Prozess (vgl. Abschnitt 2.2). Hierzu werden zuerst grundsätzliche Details von Prognosen betrachtet, d. h. welche Granularität von Prognosen möglich ist und wie diese in verfügbaren Publikationen umgesetzt werden. Nachdem die Zielidentifikation durchgeführt wurde (Was genau soll prognostiziert werden?), erfolgt die Auswahl einer Datenquelle und das Extrahieren und Einlesen dieser Daten. Im nächsten Schritt wird diese Datenmenge verarbeitet und erweitert, beispielsweise können aus den einzelnen Spielergebnissen Formtabellen aggregiert werden. Darüber hinaus wird in diesem Schritt analysiert, inwiefern die verfügbaren Attribute einen Zusammenhang zum vorherzusagenden Attribut (Spielergebnis) haben. Auf Basis dieser Korrelationsanalyse erfolgt die Auswahl von Attributen, auf welchen die verschiedenen Data-Mining-Algorithmen angewandt werden; abgeschlossen wird dies mit einer Auswertung der Ergebnisse dieser Analysen.

5.1 Ziel der Vorhersagen

Mit Prognosen im Fußball verbindet man grundsätzlich die Prognose einzelner Spielergebnisse, jedoch ist auch eine gröbere Granularität möglich, beispielsweise versucht Heuer (2013, S. 167 ff.) die Prognose einer ganzen Saison, so dass sowohl der Gewinner der Meisterschaft, als auch die möglichen Absteiger prognostiziert werden. Für die Prognosen im Rahmen dieser Arbeit wird jedoch die kleinste Granularität, nämlich das Ergebnis der einzelnen Spiele, gewählt. In diesem Fall ist nach Heuer und Rubner (2012, S. 1) eine weitere Unterteilung möglich, so kann einerseits eine ergebnisorientierte Prognose (*results-based prediction*) durchgeführt werden, bei welcher nur das Ergebnis des Spiels betrachtet wird (sog. Tendenz, mögliche Ergebnisse sind Heimsieg, Unentschieden oder Auswärtssieg); andererseits bezeichnen Heuer und Rubner (2012) die tororientierte Prognose (*goals-based prediction*) als eine detailliertere Prognose des Spielergebnisses, da hier das genaue Ergebnis durch die Prognose der Toranzahl der beiden Teams betrachtet wird, wodurch es im Gegensatz zur Prognose der Tendenz deutlich mehr Ergebnisse gibt (z. B. bei Heimsieg: 1:0, 2:1, 3:1, 3:2 usw.). In verschiedenen Forschungsarbeiten werden teilweise Zwischenstufen der beiden Ansätze verwendet, beispielsweise prognostizieren Rotshtein u. a. (2005, S. 620) die folgenden Stufen von Spielergebnissen:

- **high-score loss**: Tordifferenz = –5, –4, –3
- **low-score loss**: Tordifferenz = –2, –1
- **draw game**: Tordifferenz = 0
- **low-score win**: Tordifferenz = 1, 2
- **high-score win**: Tordifferenz = 3, 4, 5

Um den Umfang und die Komplexität diese Arbeit zu beschränken, wurde die Prognose der *Tendenz des Spielausgangs* ausgewählt, da dies das Ausgangsproblem und die Durchführung der Analysen aufgrund der Einschränkung auf nur drei mögliche Spielergebnisse etwas vereinfacht, gleichzeitig aber die Realität weiterhin korrekt beschreibt.

5.2 Forschungsstand

Während der Bearbeitung dieser Arbeit ließen sich etwa 50 Publikationen zum Themenbereich der Anwendung von Prognosetechniken im Fußball finden, wobei sich ein Teil dieser Arbeiten auf grundsätzliche sportwissenschaftliche Bereiche konzentriert, während andere konkrete Analysen auf der Grundlage der in Abschnitt 5.1 genannten Granularitätsstufen durchführen. Im deutschsprachigen Bereich ist Prof. Dr. Andreas Heuer des Instituts für Physikalische Chemie der Universität Münster recht bekannt, da dieser regelmäßig wissenschaftliche Publikationen und Artikel in populärwissenschaftlichen Zeitschriften veröffentlicht (vgl. Universität Münster, 2016). Im Vergleich zu anderen Veröffentlichungen sind die Arbeiten Heuers recht ausführlich und betrachten vor allem auch die Datengrundlage und deren Korrelation zum Erfolg einer Mannschaft (z. B. Heuer 2013, S. 67 ff.), während sich viele Veröffentlichungen auf die angewandten Algorithmen und deren theoretische Betrachtung konzentrieren (z. B. Rotshtein, Posner und Rakityanskaya 2005, S. 619 ff.).

Eine weitere Arbeit, auf die von verschiedenen Autoren (u. a. Biermann 2011, S. 181 ff.) und in verschiedenen Wettforen im Internet verwiesen wird, sind die anonym veröffentlichten Gedanken zum Thema Wettstrategien eines offenbar erfolgreichen Sportwetters (vgl. Anonym, 2005). Dieser stellt in einem etwa 25 Seiten langen Aufsatz ausführlich dar, wie er die Spiele analysiert, auf die er Wetten abschließt: hierzu werden insgesamt fünf verschiedene Bereiche betrachtet: neben grundsätzlichen Statistiken der jeweiligen Ligen (z. B. Verteilung von Heimsiegen, Unentschieden und Auswärtssiegen) werden die aufeinander treffenden Teams ausführlich betrachtet, indem unter anderem die Form der Mannschaften und die einzelnen Spieler und deren Wichtigkeit für die Mannschaften bewertet werden. Am Beispiel des Spiels *Basel – Meyrin* der zweiten Schweizer Liga stellt dieser Autor die Berechnung der Gewinnwahrscheinlichkeiten dar. Im Verlauf der Recherchen im Rahmen dieser Masterarbeit wurden neben dieser Veröffentlichung eines anonymen Autors und den Veröffentlichungen von Heuer keine vergleichbaren Publikationen gefunden, was die Ausführlichkeit der Darstellung einer Spielanalyse zur Prognose angeht.

Die folgende Darstellung ausgewählter Analysen verschiedener Autoren beschränkt sich auf die Nennung der verwendeten Verfahren und erreichten Genauigkeiten; für die Darstellung der jeweils verwendeten Attribute sei auf die Originale verweisen, da dies an dieser Stelle den Umfang sprengen würde. Die folgende Publikationen prognostizierten – wie die vorliegende Arbeit – auf Basis der Granularität Heimsieg, Unentschieden und Auswärtssieg, also ohne Betrachtung von Toren oder der Tordifferenz:

- Carpita u. a. (2015) wenden das Klassifikationsverfahren Random Forest auf der Datenbasis der italienischen Liga für die vier Saisons 2008/2009 bis 2011/2012 an und erreichen eine Genauigkeit von 64 %.
- Ulmer und Fernandez (2014) verwenden die Algorithmen Naive Bayes, Markov Model, Support Vector Machine (SVM) und Random Forest basierend auf der englischen Premier League für die Spielzeiten von 2002/2003 bis 2011/2012 und erzielen Genauigkeiten von 49 % bis zu 51 %.
- Gomes u. a. (2015) setzen Entscheidungsbäume, Naive Bayes und SVM ein und erreichen Genauigkeiten von 47 % bis 51 %; hierbei wurden die Spielzeiten 2000/2001 bis 2012/2013 der englischen Premier League betrachtet.
- Rotshtein u. a. (2005) haben durch die Anwendung von evolutionären Algorithmen und Fuzzy-Methoden auf Datensätzen der finnischen Fußballliga von 1991 bis 2001 eine Genauigkeit von etwa 80 % erreicht, wobei diese die im vorherigen Abschnitt genannte Granularität verwendet haben (high-score loss, low-score loss, etc.).
- Buursma (2015) führte Prognosen der niederländischen Fußballliga auf Datensätzen der letzten 15 Jahre durch und wandte hier Naive Bayes, bayesianische Netzwerke, logistische Regression und weitere Algorithmen an; hierbei wurden Genauigkeiten von 53 % bis 55 % erreicht.
- Owramipur u. a. (2013) haben ein bayesianisches Netzwerk auf Datensätzen der Saison 2008/2009 des spanischen Vereins FC Barcelona angewandt und eine Genauigkeit von 92 % erreicht, was auf den ersten Blick ein sehr gutes Ergebnis scheint, jedoch muss beachtet werden, dass Barcelona in dieser Saison nur fünf ihrer insgesamt 38 Spiele verloren hat (vgl. Transfermarkt.de, 2016a).

Wie aus dieser Aufstellung ersichtlich wird, schwanken die Ergebnisse deutlich von etwa 50 % bis hin zu 80 %, wobei das Extrembeispiel des FC Barcelona mit einer Genauigkeit von 92 % aufgrund der Basiswerte nicht sehr aussagekräftig ist. Je nach Liga und Wettbewerb sind unterschiedliche Ergebnisse zu erwarten, so war die in der finnischen Liga erreichte Genauigkeit von 80 % in keiner Prognose der „großen“ europäischen Ligen erreichbar; im Gegenteil, hier wurden in den Publikationen, die alle drei möglicher Ergebnisse betrachten, „nur“ Genauigkeiten im Bereich von 45-65 % erreicht.

5.3 Datengrundlage

Für die Analysen wurde aufgrund des regionalen Bezuges die deutsche Bundesliga ausgewählt, welche aus 18 Mannschaften besteht, wobei im Verlauf einer Saison alle Teams jeweils zweimal gegeneinander spielen (je ein Heimspiel und ein Auswärtsspiel). Pro Saison gibt es somit insgesamt 306 Spiele ($18 \cdot 17$ Spiele). Theoretisch wäre es möglich, sämtliche Spiele seit der Gründung der deutschen Bundesliga zu betrachten, jedoch verändern sich die Mannschaften im Laufe der Jahrzehnte, so dass Mannschaften, die früher überdurchschnittlich gut waren, heutzutage teilweise nicht mehr in der ersten Bundesliga spielen oder dort nur einen geringen Einfluss haben, während neue Mannschaften, die erst seit wenigen Jahren in der Bundesliga sind, heute sehr erfolgreich sind. Als Datengrundlage wurde daher die Betrachtung der letzten vier Spielzeiten ausgewählt, wobei die 2010/2011 bis zur Hinrunde der Saison 2015/2016 (1377 Spiele) als Trainingsdaten verwendet werden, um die 153 Spiele der Rückrunde der vergangenen Saison 2015/2016 zu prognostizieren.

Wie im Abschnitt 4.2.2 dargestellt, gibt es neben den einzelnen Spielergebnissen, aus welchen die Tabelle generiert werden kann, taktische Spieldaten (z. B. Ballkontakte, Pässe, Schüsse etc.). Da in dieser Arbeit aus naheliegenden Gründen kein kommerzieller Anbieter verwendet kann (die Preise liegen, wie dargestellt, bei mehreren Tausend Euro), muss auf frei verfügbare Fußballwebsiten zurückgegriffen werden. Diese haben jedoch, wie im Abschnitt 4.3 dargestellt, den Nachteil, dass der Umfang der Datenbestände begrenzt ist, so dass beispielsweise nur Daten auf der Granularitätsebene des gesamten Teams vorhanden sind, jedoch nicht auf der einzelner Spieler. Da eine Prognose basierend auf Statistiken einzelner Spieler jedoch auch recht komplex wäre, würde dies für diese Masterarbeit ohnehin eher nicht in Frage kommen, da hierzu Modelle erzeugt werden müssen, die darstellen, wie die Qualität einzelner Spieler Auswirkungen auf den Erfolg der ganzen Mannschaft hat, wozu auch ein sportwissenschaftlicher Hintergrund notwendig wäre.

Als Datenquelle würden sich diverse Webseiten eignen, im Bezug auf den Datenumfang stellen diese grundsätzlich dieselben Inhalte bereit. Letztlich wurde die Seite *transfermarkt.de* ausgewählt, da diese neben den Spielergebnissen, Spieldaten und taktischen Daten als einziger Anbieter auch Marktwerte der einzelnen Spieler bereitstellen. Der Marktwert eines Spielers stellt dessen Wert dar, den ein anderer Verein für diesen Spieler zu zahlen bereit wäre – derartige Zahlen werden jedoch nicht von Vereinen veröffentlicht, daher schätzt transfermarkt.de diese Werte in regelmäßigen Abständen durch die Bewertung der aktuellen Spielleistungen der Spieler (vgl. Transfermarkt.de, 2016b). In anderen Publikationen werden die Marktwerte im Übrigen ebenfalls als Datengrundlage verwendet (u.a. Heuer und Rubner, 2012, S.6).

Ein Problem, das bei der Verwendung eines kommerziellen Anbieters nicht vorhanden wäre, liegt darin, dass frei verfügbare Websites keinen direkten Datenzugriff (z. B. durch XML-Dateien oder Webservices) ermöglichen, da die Statistiken nur über die Integration in HTML-Dateien veröffentlicht werden. Zur Extraktion der Daten aus Webseiten wird

im Rahmen dieser Arbeit *jsoup* verwendet, eine Java-Library auf Open-Source-Basis, mit welcher HTML-Dateien geladen und durch die Navigation anhand des *Document Object Model (DOM)* verarbeitet werden können (vgl. Hedley, 2016). Hierbei muss der Pfad zu den HTML-Elementen angegeben werden, welche die benötigten Informationen darstelle. Nachstehend wird ein vereinfachtes Beispiel der Extraktion der Namen der Teams und des Spielergebnisses unter der Verwendung dieser Library gezeigt:

```
Document document = Jsoup.connect("http://www.transfermarkt.de/[...]").get();
MatchBean match = new MatchBean();

// allgemeine Daten:
Elements spielbericht = document.select(".spielbericht");

// Name des Heim-Teams
String heimTeamID = spielbericht.select(".heimTeam").id();
match.setHeimTeamID(heimTeamID);

// Name des Gast-Teams
String gastTeamID = spielbericht.select(".gastTeam").id();
match.setGastTeamID(gastTeamID);

// Endstand
String spielstand = spielbericht.select(".sb-endstand").text();
match.setSpielstand(spielstand);
```

Beim Anbieter Transfermarkt.de lassen sich folgende Spieldaten extrahieren:

- Datum des Spieles,
- Name des Heim-Teams,
- Name des Gast-Teams,
- Marktwert des Heim-Teams,
- Marktwert des Gast-Teams,
- Endergebnis.

Außerdem können für beide Teams jeweils folgende taktische Daten extrahiert werden:

- Anzahl der gesamten Torschüsse,
- Anzahl der Schüsse auf das Tor,
- Anzahl der Schüsse neben das Tor,
- Anzahl gehaltener Torschüsse,
- Anzahl der Eckbälle,
- Anzahl der Freistöße,
- Anzahl der Fouls,
- Anzahl der Abseits.

Theoretisch wäre außerdem eine Erfassung weiterer Statistiken möglich, unter anderem die Aufstellungen beider Mannschaften, Auswechslungen, gelbe und rote Karten, sowie Namen der Spieler mit Torvorlagen und Toren. Im Rahmen dieser Analysen wird jedoch auf die Verarbeitung derartiger Daten verzichtet, da dies die Komplexität deutlich steigern würde. So empfehlen beispielsweise Rotshtein u. a. (2005, S. 628) die Vernachlässigung folgender Kennzahlen:

- Verletzte und geschonte Spieler,
- Psychologische Faktoren,
- Schiedsrichter-Objektivität,
- Wetter und Klima.

Auf dieser erfassten Datenbasis können nun die einzelnen Schritte der Datenverarbeitung durchgeführt werden, um einen für die Datenanalysen optimierten Datenbestand zu erzeugen.

5.4 Verhinderung von Overfitting

Unter Overfitting versteht man die Überanpassung von Modellen auf die Trainingsdatenmenge (vgl. Weiss und Indurkhya, 1998, S. 43). Dies führt zu einer Verschlechterung der Qualität des Modells, da die Einordnung von neuen Datensätzen, die sicherlich nicht genau dieselben Ausprägungen wie die Trainingsdaten besitzen, ungenauer wird.

Andererseits gibt es, Weiss und Indurkhya (1998, S. 43) zufolge, jedoch auch den Fall des Underfitting, so dass das entstandene Modell sehr allgemein ist und eine geringere Anzahl an Attributen verwendet. Dies hat zwar den Nachteil, dass evtl. aussagekräftige Informationen nicht berücksichtigt werden, jedoch auch den Vorteil, dass das Modell allgemeingültiger und somit bei der Anwendung auf die Testdaten im Vergleich zum Overfitting zu einer Verbesserung der Qualität führt. Daher gilt es, einen Weg zwischen Over- und Underfitting zu finden.

Ein Beispiel für Overfitting: in der Saison 2011 ging das Spiel zwischen dem VfB Stuttgart und Schalke 04 am ersten Spieltag mit dem Ergebnis 3:0 aus, wobei der VfB Stuttgart vor dem Spiel auf Platz 12 war und innerhalb der Formtabelle durchschnittlich 1,147 Punkte pro Spiel erhielt, während Schalke 04 auf Platz 14 war und durchschnittlich 1,088 Punkte erhielt. Ein Modell könnte nun diese Attribute genau berücksichtigen, so dass eine Regel erstellt wird, die etwa folgendermaßen lauten könnte:

> IF home = Stuttgart AND away = Schalke AND spieltag = 1 AND homeTable = 12 AND awayTable = 14 AND homeFormPoints = 1,147 AND awayFormPoints = 1,088 THEN Heimsieg

Diese Modell würde nun bei der Anwendung auf Spielen außerhalb der Saison 2011 schlecht abschneiden, da diese Konstellation mit an Sicherheit grenzender Wahrscheinlichkeit nicht

vorkommen wird. Andererseits könnte nun auch folgende Regel generiert werden, die den Fall des Unterfittings darstellt:

> IF awayTable > 10 THEN Heimsieg

Eine Stufe zwischen Overfitting und Underfitting könnte beispielsweise die folgende Regel sein, deren Ergebnis vermutlich etwas besser wäre:

> IF homeTable < awayTable AND homeFormPoints > 1,100 AND awayFormPoints < 1,100 THEN Heimsieg

Daher ist es, Weiss und Indurkhya (1998, S. 83) zufolge, bei der Vorbereitung üblich, die folgenden drei Schritte zur Reduktion der Daten durchzuführen, mit dem Ziel, Overfitting zu verhindern:

- Beschränkung hinsichtlich Datensätze,
- Beschränkung hinsichtlich Attribute,
- Attributwerte vereinfachen / mögliche Werte beschränken.

In dieser Arbeit finden sich alle drei Schritte wieder. Wie bereits dargestellt, werden nur die letzten sechs Spielzeiten berücksichtigt und nicht alle Spiele seit Beginn der Bundesliga. Im Abschnitt 5.5 erfolgt die Beschränkung der zu verwendenden Attribute, im folgenden Absatz wird noch kurz auf die Vereinfachung der Attributwerte eingegangen.

Um eine Überanpassung zu vermeiden, ist es also auch sinnvoll, die möglichen Ausprägungen der Attribute einzuschränken, da dies auch zu allgemeineren Modellen führt, denn durch Zahlen mit vielen Nachkommastellen lassen sich Datensätze im Zweifelsfall eindeutig identifizieren. Dies lässt sich im Fall von kontinuierlichen Wertebereichen durch Runden umsetzen (vgl. Weiss und Indurkhya, 1998, S. 101ff.), weshalb in dieser Arbeit sämtliche Attribute der Datengrundlage vor der Verarbeitung durch die Data-Mining-Algorithmen auf eine Nachkommastelle gerundet werden.

5.5 Merkmalsauswahl

Im Fußball lassen sich unzählige Attribute erfassen und viele weitere errechnen. Um das Overfitting von Modellen verhindern, wird nur eine begrenzte Auswahl von diesen als Input für die Algorithmen verwendet werden.

In einem ersten Schritt können Attribute, mit welchen die Datensätze eindeutig identifiziert werden können, aus der Grundmenge der Attribute entfernt werden. Im vorliegenden Fall sind dies beispielsweise Datum und Uhrzeit eines Spieles, sowie die Nummer des Spieltages. Darüber hinaus lassen sich hierzu auch die Namen der beteiligten Teams zählen, jedoch enthalten diese – im Gegensatz zu Datum und Uhrzeit – auch relevante Informationen. Um allgemeingültige Modelle zu erzeugen, wird im Rahmen dieser Arbeit jedoch auf die Verwendung der Namen der Teams verzichtet, so dass lediglich numerische Kennzahlen mit Bezug auf die Qualität und Leistung der Mannschaften verwendet werden.

Im zweiten Schritt wird nun die *Korrelation* herangezogen, welche sich nach Runkler (vgl. 2015, S.59) folgendermaßen definieren lässt:

> Die Korrelation quantifiziert den Grad des Zusammenhangs zwischen Merkmalen. Ziel der Korrelationsanalyse ist es, zusammenhängende Merkmale zu identifizieren, um Ursachen für beobachtete Effekte zu erklären oder gezielt bestimmte Effekte herbeiführen zu können.

Die Berechnung der Korrelation wird durch das Programm *Excel* von Microsoft durchgeführt, welches den sog. Korrelationskoeffizienten von Bravais und Pearson implementiert (vgl. Matthäus und Schulze, 2015, S. 131). Dieser kann Matthäus und Schulze zufolge Werte zwischen -1 und $+1$ annehmen, wobei ein Wert „dem Betrag nach annähernd gleich Eins (also gleich -1 oder $+1$)“ aussagt, dass ein nahezu linearer Zusammenhang vorhanden ist; liegt der Koeffizient bei etwa Null, so ist kein Zusammenhang zwischen den betrachteten Merkmalen vorhanden.

Da es im Rahmen dieser Arbeit gilt, das Spielergebnis vorherzusagen, werden die Korrelationskoeffizienten zwischen verschiedenen Merkmalen und dem Spielergebnis berechnet, d. h. wird untersucht, welche Kennzahlen einen erhöhten Zusammenhang mit dem Erfolg einer Mannschaft besitzen. Diese Analysen werden auf dem Datenbestand der Spielzeiten von 2010/2011 bis zur Hinrunde 2015/2016 durchgeführt, da dies die Trainingsdaten für die später durchzuführenden Data-Mining-Algorithmen sind.

5.5.1 Spielbezogene Daten

In einem ersten Schritt wird untersucht, welche Kennzahlen eines Spiels eine Verbindung zum Spielergebnis besitzen, d. h. welche Kennzahlen charakteristisch dafür sind, dass eine Mannschaft ein Spiel gewinnt. Hierbei wird also jeweils der Zusammenhang zwischen Kennzahlen eines Spiels und dem Ergebnis dieses Spiels betrachtet. Damit der sog. Heimvorteil (der Vorteil der Heimmannschaft gegenüber der Gastmannschaft, beruhend auf der Tatsache, dass die Heimmannschaft statistisch gesehen eine höhere Erfolgswahrscheinlichkeit besitzt, vgl. Abschnitt 5.7.1) nicht im Korrelationskoeffizient berücksichtigt wird, werden die Daten im Folgenden so angepasst, dass nicht direkt ersichtlich ist, ob die betrachtete Mannschaft zuhause oder auswärts gespielt hat. Dies wird dadurch umgesetzt, dass aus jedem Spiel zwei Datensätze generiert werden - einer für die Statistiken der Heimmannschaft, einer für die der Gastmannschaft (daher insgesamt 2754 Datensätze für die betrachteten 1377 Spiele). Zur Betrachtung der Korrelation des Heimvorteils wird außerdem ein Attribut Heimspiel erzeugt, die aussagt, ob das Spiel der betrachteten Mannschaft ein Heimspiel war.

Kennzahl	**Korrelation**
Heimspiel (true | false)	0,172
Tore geschossen	0,668
Gegentore erhalten	–0,614
Ballbesitz	0,187
Torschüsse insgesamt	0,307
Torschüsse auf das Tor	0,399
Torschüsse neben das Tor	0,064
Gehaltene Torschüsse	–0,063
Abseits	0,037
Fouls	–0,035
Eckbälle	0,043

Tabelle 4: Korrelation zwischen den taktischen Daten und dem Spielergebnis

Wie aus der Tabelle 4 ersichtlich wird, liegen für die Kennzahlen 'Tore geschossen' und 'Gegentore gehalten' recht hohe Korrelationskoeffizienten vor. Die Erklärungen hierfür sind recht trivial: die Anzahl der geschossenen Tore sagt aus, dass Mannschaften, die in einem Spiel viele Tore schießen, relativ oft gewinnen; während der negative Wert bei den erhaltenen Gegentoren aussagt, dass Mannschaften, die wenig Gegentore erhalten, häufiger gewinnen bzw. dass Mannschaften, die viele Gegentore erhalten, deutlich seltener gewinnen. Bei den restlichen Kennzahlen stechen ansonsten lediglich die beiden Kennzahlen der gesamten Torschüsse und der Torschüsse auf das Tor heraus; während die letzten fünf Einträge aufgrund ihrer Nähe zu 0 nahezu keinen Zusammenhang zum Spielergebnis besitzen. Des Weiteren kann die Existenz des Heimvorteils bestätigt werden, da das Attribut Heimspiel eine Korrelation zum Spielergebnis besitzt, die deutlich über 0 liegt.

Eine Verwendung derartiger Statistiken wäre jedoch nicht umsetzbar, da bei den oben verwendeten Daten davon ausgegangen wurde, dass die Statistiken eines Spieles zum Zeitpunkt der Analyse bekannt sind – in der Realität hingegen würden die Prognosen vorab erstellt werden, so dass keine der oben genannten Statistiken vorhanden ist. Nichtsdestotrotz wurden in diesem Schritt die Kennzahlen erfolgreicher Mannschaften herausgefunden, auf deren Basis im nächsten Abschnitt weitere Untersuchungen durchgeführt werden, indem die Daten vergangener Spiele zur Prognose herangezogen werden.

Der nächste logische Schritt ist nun die Betrachtung der sog. *Form* der aufeinandertreffenden Mannschaften, die darstellt, wie erfolgreich eine Mannschaft in den letzten n Spielen war, wobei untersucht werden muss, welche Spielanzahl zur Berechnung der Form am aussagekräftigsten ist. Zuvor erfolgt jedoch noch die Betrachtung statischer Daten.

5.5.2 Statische Daten

Unter statischen Daten versteht man Informationen, die keinen direkten Bezug zu den taktischen Daten eines Spiels, wie Ballbesitz, Torschüsse etc., haben und somit gleichsam fix definiert sind. Hierzu gehören die aktuellen Positionen der beiden Mannschaften in der Tabelle, die Positionen in der Abschlusstabelle der letzten Saison und der erfasste Marktwert, da dieser nur am Anfang einer Saison und zur Saisonmitte erfasst wird. Im Bezug auf die Abschlussposition der letzten Saison lässt sich neben den Positionen der Mannschaften die Differenz der Positionen betrachten. Ein Problem hierbei ist, dass die Abschlusspositionen neuer Mannschaften, die aus der zweiten Liga aufgestiegen sind, nicht definiert sind. Ein möglicher Ansatz wäre, die Abschlussposition dieser Mannschaften aus der zweiten Liga zu nehmen, was aber nicht praktikabel wäre, da diese Mannschaften dann die Abschlusspositionen 1-3 hätten. Heuer (2013, S. 104 f.) zufolge stieg seit 1995 jeder dritte Aufsteiger sofort wieder ab, wobei die Abschlusspositionen in nahezu sämtlichen Fällen in der unteren Tabellenhälfte (Plätze 10-18) lagen. Heuer empfiehlt, durchschnittliche Werte für die unbekannten Datensätze der Aufsteiger zu verwenden, aus diesem Grund wird im Folgenden die Abschlussposition 14 verwendet (Mittelwert aus 10 und 18).

Ein weiteres Problem liegt darin, dass die Marktwerte der Teams in den letzten Jahren aufgrund der erhöhten Kommerzialität des Fußball stiegen, so stieg die Summe der Marktwerte aller 18 Mannschaften der Bundesliga im betrachteten Zeitraum von 1,7 Mrd. Euro in der Saison 2010/2011 zu 2,4 Mrd. Euro in der Saison 2014/2015. Im Folgenden wird der normale Marktwert betrachtet, bei den Analysen hat sich jedoch herausgestellt, dass sich eine saisonübergreifend standardisierte Form besser eignet. Hierzu wird das Team mit dem niedrigsten Marktwert mit 0 und das Team mit dem höchsten Marktwert mit 1 bewertet; für alle anderen Teams erfolgt die Berechnung des Wertes durch eine lineare Interpolation zwischen diesen Extremen. Die Korrelationskoeffizienten des standardisierten Marktwertes besitzen nur minimale Unterschiede zu den normalen Marktwerten.

Wie aus Tabelle 5 ersichtlich wird, fallen die Korrelationskoeffizienten im Vergleich mit dem vorherigen Abschnitt deutlich niedriger aus. Für die aktuelle Tabellenposition wurde ein Wert von $-0{,}166$ ermittelt, d. h. je niedriger die Tabellenposition ist, desto größer ist die Wahrscheinlichkeit für einen Erfolg einer Mannschaft; bei der Betrachtung der gegnerischen Mannschaft hingegen ist ein positiver Koeffizient vorhanden, d. h. je schlechter die Tabellenposition des Gegners ist, desto wahrscheinlicher ist ein Erfolg der betrachteten Mannschaft. Die Korrelation mit der Abschlussposition lässt sich verbessern, indem die Differenz zwischen den Positionen der betrachteten Mannschaften errechnet wird. Spielt beispielsweise der Tabellenerste gegen den Tabellenletzten, so wäre diese Differenz bei -17; spielt hingegen der letzte gegen den ersten, so läge dieser Wert bei $+17$.

Analog gilt dies auch für die Abschlusspositionen der Tabelle der letzten Saison, hier sind die Korrelationskoeffizienten jedoch leicht höher.

Kennzahl	**Korrelation**
Aktueller Tabellenplatz Heim	–0,166
Aktueller Tabellenplatz Gast	0,160
Differenz Tabellenplatz	–0,226
Abschlussposition Vorjahr Heim	–0,173
Abschlussposition Vorjahr Gast	0,160
Differenz Abschlussposition	–0,237
Marktwert Heim	0,239
Marktwert Gast	–0,216
Differenz Marktwert	0,315
Marktwert Heim standardisiert	0,236
Marktwert Gast standardisiert	–0,216
Differenz Marktwert standardisiert	0,312

Tabelle 5: Korrelation zwischen statischen Daten und dem Spielergebnis

5.5.3 Gemittelte Daten / Formdaten

Um sinnvolle Vorhersagen treffen zu können, müssen Kennzahlen vorhanden sein, die ausdrücken, welche Qualität eine Mannschaft besitzt. Dies sollte jedoch keine statische Kennzahl sein, da die Qualität innerhalb einer Saison nicht gleich bleibt. Daher wird hierzu die aktuelle *Form* einer Mannschaft hinzugezogen, d. h. die Ergebnisse der letzten Spiele vor dem untersuchten Spiel. Diese Kennzahl kann nun an jedem Spieltag für die letzten n Spiele errechnet werden, wobei der zu betrachtende Zeitraum der Untersuchungsgegenstand dieses Abschnitts ist. In einem ersten Schritt kann dieser Zeitraum auf ein einzelnes Spiel beschränkt werden, so dass nur das letzte Spiel vor dem aktuellen Spiel betrachtet wird, die entsprechenden Korrelationskoeffizienten hierzu befinden sich in Tabelle 6. Für das erste Saisonspiel wäre hier kein letztes Spiel vorhanden, daher wird hier das Spiel des letzten Spieltages der vorherigen Saison verwendet. Dies führt jedoch dazu, dass für die Aufsteiger kein letztes Spiel vorhanden ist, da diese in der vorherigen Saison noch in der zweiten Liga waren. Für Aufsteiger werden diese Datensätze daher nicht erzeugt (11 Datensätze für die betrachteten sechs Saisons).

Kennzahl	**Korrelation**
Erzielte Punkte	0,029
Tore geschossen	0,063
Gegentore erhalten	–0,018
Torschüsse gesamt	0,072
Torschüsse auf das Tor	0,064

Tabelle 6: Korrelation zwischen Daten des letzten Spiels und dem Spielergebnis

Die hier errechneten Werte liegen nahezu bei 0, was bedeutet, dass kein Zusammenhang zwischen dem Erfolg einer Mannschaft im betrachteten Spiel und dem Erfolg im vorherigen Spiel vorhanden ist. Begründen lässt sich dies damit, dass der Zufall bei einem so eng gesteckten Zeitraum eine recht große Rolle besitzt; außerdem kann es natürlich sein, dass das vorherige Spiel gegen einen überdurchschnittlich guten oder einen unterdurchschnittlich schlechten Gegner war, wodurch dies nur eine geringe Aussagekraft im Bezug auf die Qualität der Mannschaft besitzt.

Eine akkuratere Abbildung der Form erfordert daher die Betrachtung mehrerer Spiele, beispielsweise betrachtet die offizielle Formtabelle (vgl. Deutsche Fußball Liga GmbH, 2015) die letzten fünf Spiele; der oben bereits verwiesene Wettanalyst (vgl. Anonym, 2005) verwendet für die Analyse der Form ebenfalls die letzten fünf Spiele, wobei dieser die einzelnen Spiele grundsätzlich mit je 20 % gewichtet, korrigiert um eine subjektive Einschätzung. Heuer und Rubner (2013, S. 6) hingegen betrachten im Rahmen der Form die letzten sieben Spiele. Im Folgenden wird die Form aufgrund des Fehlens einer einheitlichen Definition nicht fix definiert, sondern für verschiedene Werte hinsichtlich ihrer Korrelation zum Spielerfolg untersucht.

Würde man nun z. B. die letzten fünf Spiele betrachten, sind am Anfang einer Saison, also z. B. am ersten Spieltag, noch keine Daten vorhanden. Daher wird die Form einer Mannschaft im Folgenden saisonübergreifend errechnet, d. h. es besteht die Form einer Mannschaft am ersten Spieltag aus den letzten fünf Spielen der letzten Saison. Wie oben besteht hier ebenfalls das Problem, dass für Aufsteiger keine Daten der Vorsaison vorhanden sind. Während es oben lediglich 11 Spiele waren, für die keine Datensätze generiert werden können, betrifft dies hier deutlich mehr Spiele, insbesondere für den Fall, dass die Form über einen längeren Zeitraum als fünf Spiele betrachtet wird (bei $n = 15$ wären es z. B. über 300 Spiele, da an den ersten 15 Spieltagen jeweils sämtliche Spiele der Aufsteiger ignoriert werden müssten). Daher wird die Form für Aufsteiger geändert: während die Form „normaler“ Teams z. B. immer aus 15 Spielen besteht, muss die Form eines Aufsteigers aus mindestens drei Spielen errechnet werden (der Wert von drei Spielen wurde anhand der Berechnung aller Korrelationen für die Anzahlen von 1–5 gewählt). Daher kann die Formtabelle ab dem vierten Spieltag für sämtliche Mannschaften errechnet

werden, so dass für die gesamte Trainingsdatenmenge lediglich 32 Spiele ignoriert werden. Da hier jedoch – im Gegensatz zur alleinigen Betrachtung des letzten Spieles – aufsummierte Daten vorhanden sind, die sich über verschiedene Zeiträume erstrecken (z. B. hat eine „normale“ Mannschaft über den Zeitraum von z. B. 15 Spielen, 30 Punkte erzielt, während ein Aufsteiger in seinen ersten drei Spielen z. B. 6 Punkte erreicht hat), werden bei der Berechnung durchgehend Durchschnitts-Werte (z. B. der Punkteschnitt) verwendet.

Kennzahl	**n=5**	**n=10**	**n=15**	**n=20**	**n=25**	**n=30**
Platz Formtabelle	–0,142	–0,187	–0,188	–0,220	–0,214	**–0,238**
Platz Formtabelle Gegner	0,131	0,171	0,171	0,205	0,199	**0,223**
Differenz Tabellenplatz	–0,195	–0,258	–0,256	–0,296	–0,289	**–0,320**
Ø Tore	0,185	0,202	0,213	0,231	0,243	**0,249**
Ø Tore Gegner	–0,171	–0,185	–0,199	–0,213	–0,222	**–0,228**
Differenz Ø Tore	0,254	0,277	0,294	0,313	0,328	**0,336**
Ø Gegentore	–0,111	–0,155	–0,180	–0,197	–0,192	**–0,204**
Ø Gegentore Gegner	0,103	0,140	0,162	0,179	0,176	**0,189**
Differenz Ø Gegentore	–0,153	–0,212	–0,248	–0,271	–0,264	**–0,281**
Ø Punkte	0,169	0,206	0,218	0,238	0,241	**0,254**
Ø Punkte Gegner	–0,154	-0,187	–0,201	–0,218	–0,222	**–0,235**
Differenz Ø Punkte	0,232	0,286	0,302	0,323	0,327	**0,343**

Tabelle 7: Korrelation zwischen Daten der Formtabelle und dem Spielergebnis

Wie in den Ergebnissen in Tabelle 7 ersichtlich ist, finden sich hier sämtliche Spalten der normalen Fußballtabelle wieder, d. h. Tabellenpositionen, geschossene Tore und erhaltene Gegentore sowie die erzielten Punkte. Diese Kennzahlen wurden außerdem miteinander verrechnet, so dass beispielsweise eine Differenz der Tabellenpositionen beider Mannschaften in der Formtabelle berechnet wurde.

Der höchste Korrelationskoeffizient wurde bei der Differenz der durchschnittlich erreichten Punktezahl erreicht, mit anderen Worten wie viele Punkte eine Mannschaft durchschnittlich mehr als die gegnerische Mannschaft erspielt hat.

Die Korrelationskoeffizienten spielbezogener Daten (z. B. Torschüsse, Ballbesitz etc.) fielen bei der Betrachtung über die Form recht gering aus (etwa 0,1), so dass auf eine Darstellung dieser verzichtet wurde, da diese im Vergleich zu den anderen Kennzahlen zu niedrig für eine etwaige Verwendung wäre.

Die Berechnung der Formtabelle (saisonübergreifend) anhand der letzten 30 Spiele besitzt durchgehend die höchsten Korrelationskoeffizienten. Bei der Betrachtung der Korrelationskoeffizienten fällt zudem auf, dass diese mit der steigenden Formlänge stetig stei-

gen, jedoch wird die Länge hier auf 30 Spiele beschränkt, da sonst Spiele der vorletzten Saison zur Form hinzugezählt werden würden, so dass die Anzahl nicht berücksichtigter Spiele steigt, da diese neben den aktuellen Aufsteigern auch für die Aufsteiger der letzten Saison anfallen würden.

5.5.4 Zusammenfassung

Folgende Kennzahlen werden aufgrund ihrer errechneten Korrelationskoeffizienten bei der Anwendung von Data-Mining-Algorithmen zur Prognose verwendet:

- Tabellenplatz in der regulären Tabelle,
- Tabellenplatz in der Tabelle der letzten Saison,
- Marktwert der Teams in der standardisierten Form,
- Daten der Formtabelle mit $n = 30$:
  - Tabellenplatz,
  - durchschnittlich erzielte Tore,
  - durchschnittlich erhaltene Gegentore,
  - durchschnittliche Punktezahl.

Um die Anzahl der Attribute klein zu halten, wurde jeweils die Kennzahl der Heimmannschaft (z. B. standardisierter Marktwert) sowie die dazugehörige Differenz verwendet (z. B. Differenz der standardisierten Marktwerte), während die Kennzahl der Gastmannschaft nicht verwendet wird. Der Grund hierfür liegt darin, dass die alleinige Verwendung der Differenzen nicht ausreicht, da ein weiterer Anhaltspunkt von Nöten ist. Beispielsweise lässt sich aus der Differenz der Marktwerte von z. B. 10 Mio. Euro nicht schließen, ob die beiden Mannschaften 15 bzw. 25 Mio., oder aber 200 bzw. 210 Mio. Euro wert sind. Hierzu sind jedoch nicht beide Kennzahlen notwendig, so dass nur die jeweilige Kennzahl der Heimmannschaft verwendet wird.

5.6 Vorgehensweise

Für die Durchführung der Analysen wird die Software RapidMiner verwendet, welche seit 2001 – anfangs unter dem Namen *Yet Another Learning Environment* (YALE) – vom Lehrstuhl für künstliche Intelligenz der Universität München entwickelt wird (vgl. Rapid-I, 2010, S. 19). Eigenen Aussagen zufolge besitzt die RapidMiner-Umgebung mehr als 1500 Operatoren, wobei sich diese auf die verschiedenen Prozessschritte verteilen, da neben den eigentlichen Algorithmen eine Unterstützung des gesamten Data-Mining-Prozesses vorgesehen ist, weshalb u. a. Operatoren für das Einlesen von Datensätzen und für deren Vorbereitung im Hinblick auf die Anwendung von Algorithmen implementiert sind (vgl. RapidMiner, 2014, S. 19).

Wie bereits dargestellt, werden nun Analysen auf Daten der deutschen Bundesliga durchgeführt, wobei die fünf Saisons 2010–2014 und die Hinrunde der Saison 2015 zur Gene-

rierung von Modellen verwendet wird, mit welchen die Spiele der Rückrunde 2015 vorhergesagt werden sollen. Die verwendeten Daten lassen sich somit zwei verschiedenen Kategorien zuordnen: *Trainingsdaten*, mit welchen die verschiedenen Algorithmen Modelle erzeugen und *Testdaten*, auf welche die erzeugten Modelle dann angewandt werden (vgl. Witten und Frank, 2001, S. 128).

Der Grund für diese Auswahl liegt darin, dass in der Rückrunde für sämtliche Spiele Daten vorhanden sind, während in der Hinrunde Datensätze für Aufsteiger fehlen können.

Pro Saison gibt es grundsätzlich 306 Spiele, wobei die verwendeten Kennzahlen nicht für alle Datensätze vorhanden sind, da die Form für neue Mannschaften an den ersten drei Spieltagen noch nicht errechnet werden kann. Daher bestehen die Trainingsdaten aus 1645 Datensätzen (statt 1683), während für die Testdaten auf alle 153 Datensätze der Rückrunde zurückgegriffen werden kann.

Die Qualität eines Modells wird hierbei durch die Fehlerrate definiert: liegt ein Modell mit seiner Prognose richtig, wird dies als Erfolg gewertet, liegt es falsch, liegt ein Fehler vor (vgl. Witten und Frank, 2001, S. 128). Die Fehlerrate gibt Witten und Frank zufolge daher das Verhältnis von Fehlern zur Anzahl der Menge der betrachteten Datensätze an. Im Folgenden wird jedoch die *Erfolgsrate* verwendet, welche sich entsprechend aus dem Anteil der Erfolge an der Gesamtzahl der Datensätze errechnet.

Eine naheliegende Vorgehensweise für eine Prognose wäre, ein für die Gesamtheit der Trainingsdaten optimales Modell zu finden und dieses auf den Testdaten anzuwenden, in der Hoffnung, dass dieses Modell eine gute Genauigkeit für diese, d. h. die Spiele der Rückrunde 2015, besitzt. In der Praxis ist es jedoch so, dass ein für die Trainingsdaten optimales Modell keineswegs optimal für die Testdaten ist. Ganz im Gegenteil: ein für die Trainingsdaten optimales Modell könnte zu sehr auf die Eigenheiten dieser Datenmenge angepasst sein.

Ein Modell zu generieren, das auf Basis der *Testdaten* optimiert wird (d h. ein Modell, dass die Saison 2015 optimal abbildet), würde logischerweise eine sehr hohe Genauigkeit bringen, jedoch entspräche dies nicht der üblichen Vorgehensweise für eine Prognose, da Testdaten nicht für die Erzeugung eines Modells verwendet werden sollten. Zudem würde ein Modell, das für die Testdaten optimal ist, aus den oben genannten Gründen nicht optimal für die Trainingsdaten sein, weshalb es – unter der Annahme, dass die Testdaten nicht bekannt sind – auf rationaler Ebene nie für eine Prognose herangezogen werden würde.

Eine Vorgehensweise, die sich aus den beiden genannten Ansätzen zusammensetzt, besteht darin, die verfügbaren Daten in insgesamt drei Teile zu unterteilen: neben den bereits genannten Testdaten, deren Klassifikation prognostiziert werden soll, werden Datensätze von den Trainingsdaten abgespalten und der Menge der Auswertungsdaten zugewiesen (vgl. Witten und Frank, 2001, S. 129 f.). Die Trainingsdaten werden nun zur Modellerzeugung

verwendet, um die Qualität dieses Modells anschließend anhand dieser Auswertungsdaten herauszufinden. Da die Klassifikation, d. h. die Spielergebnisse, bereits zum Zeitpunkt der Auswertung bekannt sind, lässt sich dieser Vorgang optimieren, so dass aus der Menge der erzeugten Modelle das beste Modell für die Auswertungsdaten ausgewählt werden kann (siehe Abbildung 6). Dieser Prozess führt somit intern bereits eine Prognose inkl. Auswertung durch, in der Hoffnung, dass das für diese interne Prognose optimale Modell auch eine gute Genauigkeit für die Testdaten besitzt.

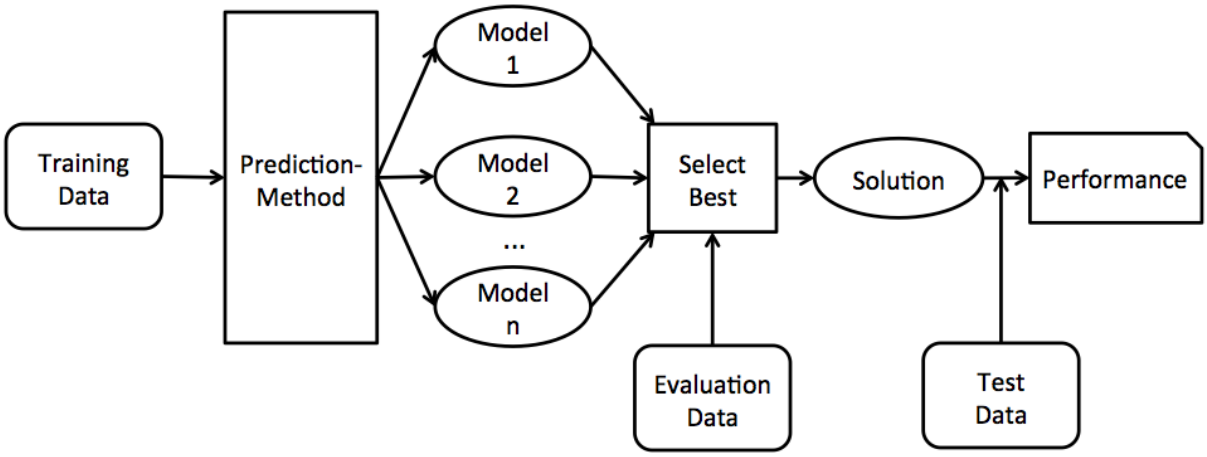


Abbildung 6: Aufbau der grundsätzlichen Vorgehensweise (in Anlehnung an Weiss und Indurkhya, 1998, S. 37)

Die Unterteilung der Trainings- und Auswertungsdaten kann nun einerseits fest definiert sein (z. B. Prognose von 2014 anhand von 2012 und 2013); andererseits bietet sich hier jedoch auch der Einsatz einer sog. Kreuzvalidierung (*Cross-Validation*) an. Hierbei findet in einer definierbaren Anzahl von Durchläufen eine Unterteilung der Trainingsdaten in mehrere Teile statt, auf deren Basis dann ein Modell erzeugt und angewandt wird (vgl. Witten und Frank, 2001, S. 133 f.). Im Rahmen dieser Arbeit erfolgte eine 10-fache Kreuzvalidierung, d. h. die Trainingsdatenmenge wurde in zehn verschiedene Teilmengen unterteilt, wobei in zehn Durchgängen dann jeweils neun Teilmengen zur Prognose der letzten Teilmenge verwendet wird.

Für die Unterteilung der Daten innerhalb der Kreuzvalidierung kann hierbei – neben einer linearen Unterteilung – auch eine zufällige Erstellung der Teilmengen durchgeführt werden; außerdem bietet das verwendete Tool *RapidMiner* die sog. stratifizierte Auswahl an (engl.: *stratified sampling*), bei welcher die Datenmenge in Teilmengen unterteilt wird, die eine ähnliche Verteilung der Ergebnisse wie die originale Gesamttestdatenmenge besitzen (vgl. RapidMiner, 2016c).

Die Verwendung einer zufälligen Auswahl hat jedoch, Weiss und Indurkhya (1998, S. 42) zufolge, den Nachteil, dass zufällig sehr einheitliche Datensätze ausgewählt werden könnten (z. B. fast nur Spiele, in denen die Heimmannschaft gewinnt), wodurch das Modell eine recht gute Genauigkeit erzielt, jedoch die Realität nur beschränkt darstellt.

Für eine lineare Auswahl spricht, dass es im Fußball die statistische Eigenheit gibt, dass es gegen Ende einer Saison tendenziell weniger Unentschieden gibt, da Mannschaften z. B. noch Punkte benötigen, um den Abstieg zu verhindern, während die Mannschaften am Anfang einer Saison noch mit einem Unentschieden zufrieden gewesen sind. Eine lineare Unterteilung auf der Ebene ganzer Saisons wäre daher sinnvoll, so dass die statistischen Eigenheiten einer Saison komplett berücksichtigt werden können. Jedoch ist dies aufgrund dessen, dass fünfeinhalb Spielzeiten betrachtet werden und dass die Anzahl der Spiele innerhalb dieser nicht gleichbleibend ist (da Spiele der Aufsteiger ignoriert werden, wobei die Anzahl der Aufsteiger nicht immer identisch ist) nicht möglich, da das Tool RapidMiner nur eine Unterteilung in gleich große Teilmengen zulässt. Daher erscheint es im vorliegenden Fall sinnvoller, statt der zufälligen oder linearen Unterteilung die Möglichkeit eines *stratified samplings* zu verwenden.

Wie bereits im Abschnitt 5.4 dargestellt, erfolgt üblicherweise eine Einschränkung auf eine Teilmenge der verfügbaren Attribute, um Overfitting zu verhindern. Hierzu wurde im vorherigen Abschnitt eine grundsätzliche Einschränkung anhand der Korrelation durchgeführt, so dass nun lediglich 15 Attribute in Frage kommen. In ersten Tests wurde nun mit diesen Attributen „experimentiert“, wobei dies unabhängig von den Testdaten durchgeführt wurde, so dass nur die Trainingsdatenmenge verwendet wurde (d. h. ohne Betrachtung der vorherzusagenden Rückrunde 2015). Hierbei hat diese große Anzahl jedoch zu Problemen hinsichtlich von Overfitting geführt, so dass sehr schlechte Genauigkeiten vorhanden waren. Aus diesem Grund bietet das Tool *RapidMiner* eine Optimierung hinsichtlich der verwendeten Attribute an: es kann eine Anzahl von Attributen angegeben werden, anhand derer alle möglichen Kombinationen der Attribute erzeugt und zur Durchführung der Algorithmen verwendet werden (vgl. RapidMiner, 2016b). Bei Tests auf der Trainingsdatenmenge hat sich hier herausgestellt, dass der Bereich 1 bis 3 für die Anzahl der Attribute am besten ist; hierbei werden alle Attribute einzeln, sowie alle 2-er und 3-er-Kombinationen verwendet. Bei der Anzahl von 15 Attributen führt dies dazu, dass es etwa 450 verschiedene Kombinationen gibt, für die jeweils durch die Kreuzvalidierung verschiedene Modelle erzeugt und angewandt werden, so dass eine Optimierung hinsichtlich deren Qualität durchgeführt werden kann. Aufgrund der 10fachen Kreuzvalidierung werden pro Algorithmus daher etwa 4.500 verschiedene Modelle erzeugt (vgl. Abbildung 6 mit $n \approx 4.500$).

5.7 Analysen

In den folgenden Abschnitten erfolgt die Darstellung der Analyseergebnisse, hierbei wird zudem jeweils eine Darstellung der theoretischen Grundlage zu den jeweiligen Algorithmen durchgeführt. Zuvor erfolgt die Betrachtung trivialer Prognosen zur Findung einer *Baseline*, um die Qualität der anderen Modelle besser einschätzen zu können.

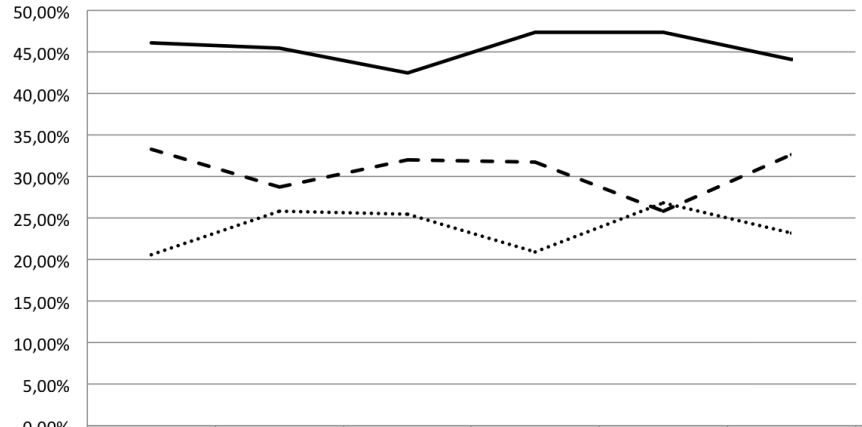
Neben den trivialen Prognosen werden folgende Algorithmen angewandt:

- Entscheidungsbäume,
- Naive Bayes,
- Künstliche neuronale Netze,
- Logistische Regression.

5.7.1 Triviale Prognosen

Aufgrund der Beschränkung der Betrachtung auf die Tendenz der Spiele gibt es nur drei mögliche Spielergebnisse: Sieg der Heimmannschaft, Unentschieden und Sieg der Gastmannschaft. Durch die Prognose anhand einer zufälligen Auswahl des Ergebnisses lässt sich daher – allerdings unter der Annahme, dass die drei möglichen Ergebnisse mit derselben Wahrscheinlichkeit ausgewählt werden – eine Genauigkeit von 33,3 % erreichen.

Durch die Betrachtung der statistischen Verteilung der drei möglichen Ausgänge lässt sich diese Genauigkeit verbessern. Hierzu befindet sich in Abbildung 7 die Darstellung dieser Verteilung, wobei die Datengrundlage die im Rahmen dieser Arbeit verwendete ist, d. h. die im vorherigen Kapitel ausgeschlossenen Spiele der Aufsteiger sind hierbei nicht miteinbezogen. Im Durchschnitt liegt der Anteil an Heimsiegen bei etwa 45 % und ist somit das häufigste Ergebnis, während Unentschieden mit rund 24 % der Ergebnisse die geringste Häufigkeit haben. Im Verlauf der letzten Jahre gab es hier zwar Schwankungen der einzelnen Häufigkeiten, im Großen und Ganzen blieb die Verteilung jedoch gleich.



	2010	2011	2012	2013	2014	2015	Durchschnitt
Anteil Sieg Heimmannschaft	46,08%	45,42%	42,48%	47,39%	47,39%	44,12%	45,48%
Anteil Unentschieden	20,59%	25,82%	25,49%	20,92%	26,80%	23,20%	23,80%
Anteil Sieg Gastmannschaft	33,33%	28,76%	32,03%	31,70%	25,82%	32,68%	30,72%

Abbildung 7: Verteilung der Ergebnisse in den Spielzeiten 2010–2015

Diese Verteilung lässt sich durch den sog. Heimvorteil erklären, den die Heimmannschaft gegenüber der Gastmannschaft besitzt. Dieser lässt sich u. a. dadurch erklären, dass die Heimmannschaft einen kürzeren Anreiseweg zum Spielort hat und eine größere Anzahl an Fans besitzt, außerdem existieren auch psychologische Faktoren, da sich die Spieler in einer gewohnten Umgebung befinden (vgl. Carron u. a., 2005, S. 396 ff.).

Für die zu prognostizierende Rückrunde 2015 liegt der Anteil an Heimsiegen bei 45,10 %, Auswärtssiege traten mit der Häufigkeit 30,07 % auf, während Unentschieden mit 24,94 % das seltenste Ergebnis sind (Anmerkung: in der obigen Abbildung ist dies nicht ablesbar, da dort die Saison 2015 als Ganzes dargestellt wird). Nimmt man nun zur Prognose ein Modell, das für jedes Spiel unabhängig von dessen Attributsausprägungen immer das häufigste Ereignis, d. h. einen Heimsieg prognostiziert, erhält man für die Rückrunde der Saison 2015 eine Genauigkeit von 45,10 %. In den folgenden Abschnitten ist es daher das Ziel, diese Genauigkeit durch die Anwendung verschiedener Algorithmen des Data Minings zu verbessern.

5.7.2 Entscheidungsbäume

Beschreibung des Algorithmus

Zur Erläuterung von Entscheidungsbäumen befindet sich in Abbildung 8 ein Beispiel, mit welchem Lebewesen anhand zwei ihrer Eigenschaften klassifiziert werden können. Am ersten Knoten wird überprüft, ob das Lebewesen größer oder kleiner als 1 Meter ist; bei den Knoten der zweiten Ebene sind Regeln im Bezug auf die Anzahl der Beine vorhanden. Durch das Durchlaufen des Entscheidungsbaumes kann somit ein unbekanntes Lebewesen anhand seiner Eigenschaften klassifiziert werden.

Unabhängig von diesem Beispiel besteht der Graph eines Entscheidungsbaums aus einem Wurzelknoten, beliebig vielen internen Knoten und Blattknoten auf der letzten Ebene (vgl. Tan u. a., 2006, S. 150). An den Wurzelknoten und den internen Knoten findet jeweils eine Prüfung auf Regeln statt, während die auf der untersten Ebene stehenden Blattknoten die verschiedenen Gruppen darstellen.

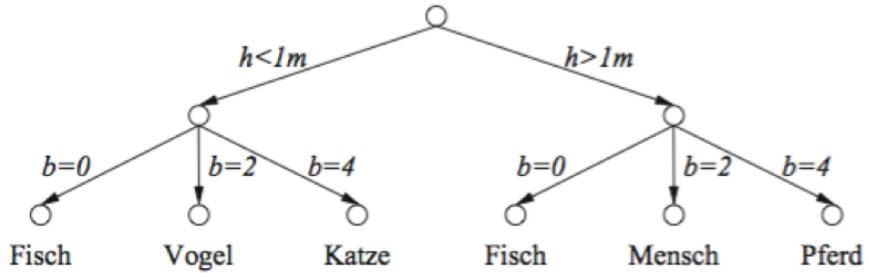


Abbildung 8: Beispiel eines Entscheidungsbaums (Runkler, 2015, S. 103)

Bei der Generierung eines Entscheidungsbaums werden in der Trainingsphase die Objekte mit bekannter Zuordnung anhand der Ausprägungen ihrer Attribute in disjunkte Teilmengen unterteilt, wodurch die Regeln an den Knoten entstehen (vgl. Cleve und Lämmel,

2014, S. 91). Im ersten Schritt muss daher ein Attribut ausgewählt, anhand dessen die erste Abzweigung durchgeführt wird. Zur Auswahl eines Attributes gibt es verschiedene Möglichkeiten, eine Möglichkeit ist nach Cleve und Lämmel (vgl. 2014, S. 95), dass jedes Attribut hinsichtlich der Genauigkeit der Vorhersage der Klassifikation untersucht wird, wobei im Anschluss das Attribut mit der höchsten Genauigkeit ausgewählt wird. Alternativ ist auch die Berechnung des Gini-Indexes oder der Entropie möglich (vgl. Cleve und Lämmel, 2014, S. 104 ff.).

Nachdem das erste Attribut ausgewählt ist, erfolgt ein Aufbau eines Teilbaums an den beiden erzeugten Knoten mit den jeweiligen Teilmengen, was rekursiv so lange wiederholt wird, bis alle Objekte einer Teilmenge derselben Klassifikation angehören (vgl. Cleve und Lämmel, 2014, 92 f.).

Hierbei ist jedoch auch das Problem des Overfittings vorhanden, da der Baum grundsätzlich beliebig groß werden kann, so dass im Extremfall alle Trainingsdaten korrekt klassifiziert werden können, die Blattknoten jeweils jedoch nur einzelne Datensätze abbilden. Mögliche Lösungsansätze hierfür sind (vgl. Cleve und Lämmel, 2014, S. 108):

- Beschränkung der Höhe des Baumes, so dass der rekursive Aufbau nach Erreichen einer bestimmten Höhe abgebrochen wird,
- Beschränkung hinsichtlich einer Mindestanzahl an Datensätzen, die jeder Blattknoten beinhalten muss,
- Entscheidungsbaum komplett generieren und im Nachhinein Teilbäume durch Blattknoten ersetzen (*pruning*); alternativ kann dies auch schon während der Erzeugung des Baumes durchgeführt werden (*pre-pruning*).

Da es sinnvoll ist, ein Overfitting zu verhindern, sollte mindestens einer der genannten Lösungsansätze verwendet werden. In Tests auf der Trainingsdatenmenge wurden Parameter für den Algorithmus erarbeitet, die zu guten Ergebnissen führen. Hierbei kann jedoch keine Optimalität sichergestellt werden, da die Trainingsdaten nicht zwingend mit den Auswertungsdaten identisch sind. Im Folgenden werden die verwendeten Parameter im *RapidMiner* aufgeführt, wobei jedoch nur jene Parameter genannt werden, bei welchen von den Standard-Werten abgewichen wurde:

- **maximale Baumhöhe**: 4
- **Pruning / Pre-Pruning**: deaktiviert
- **Kriterium zur Auswahl der Split-Attribute**: Genauigkeit (*Accuracy*)

Im übrigen gibt es für die Erzeugung eines Entscheidungsbaums verschiedene Algorithmen (u. a. ID3, CART, CHAID, vgl. Runkler, 2015, S. 105f.), auf deren Erläuterung an dieser Stelle verzichtet wird, da diese grundsätzlich nach der oben dargestellten Vorgehensweise vorgehen.

Analyseergebnisse

In Abbildung 9 wird der entstandene Entscheidungsbaum dargestellt. Dieser verwendet drei Attribute: die Differenz der Marktwerte der beiden Teams, den Tabellenplatz der Heimmannschaft in der Formtabelle sowie den Punkteschnitt der Heimmannschaft in der Formtabelle.

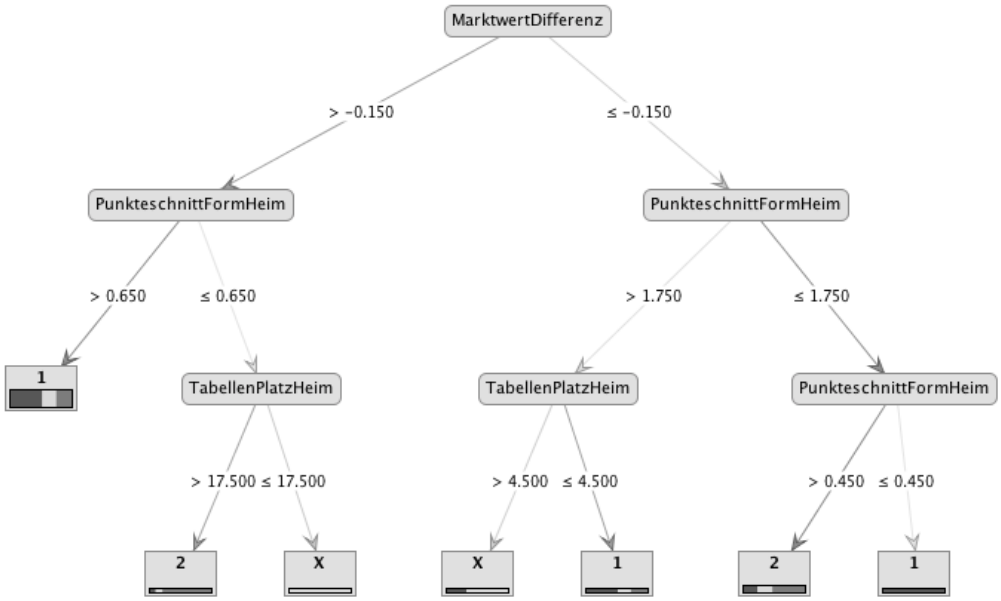


Abbildung 9: Ergebnismodell: Entscheidungsbaum

Auf der ersten Ebene, d. h. am Wurzelknoten, findet eine Überprüfung hinsichtlich der Marktwertdifferenz statt. Die Marktwerte sind standardisiert auf einen Wertebereich von 0 bis 1. Daher hat das Attribut Marktwertdifferenz einen Wert von 0, falls beide Mannschaften denselben Marktwert haben; einen Wert von -1, falls die Heimmannschaft den geringsten und die Gastmannschaft den höchsten Marktwert besitzt und einen Wert von $+1$, falls die Heimmannschaft den höchsten und die Gastmannschaft den geringsten Marktwert hat. Am ersten Knoten wird also überprüft, ob die Heimmannschaft einen deutlich geringeren Marktwert (d. h. < 0.150) als die Gastmannschaft besitzt, oder ob der Marktwert ähnlich oder größer ist.

An den weiteren Knoten findet eine Prüfung der Attribute Punkteschnitt und Tabellenplatz statt, was an dieser Stelle nicht weiter erläutert wird, da der weitere Aufbau des Entscheidungsbaums und die Bedeutung dieser Attribute selbsterklärend sind.

Die Genauigkeit der Prognosen unter Verwendung des oben abgebildeten Entscheidungsbaums beträgt 52,29 %. Zur detaillierteren Betrachtung der Analyseergebnisse wird in Tabelle 8 die Wahrheitsmatrix (engl. *Confusion-Matrix*) abgebildet, in welcher die Gesamt-Wahrscheinlichkeit auf die einzelnen Ergebnisse heruntergebrochen wird (vgl. Bramer, 2007, S. 89). In dieser wird – wie im Entscheidungsbaum oben – ein Sieg der Heimmannschaft durch eine '1' dargestellt, ein Unentschieden mit einem 'X' und ein Sieg der Gastmannschaft mit einer '2'.

		tatsächliches Ergebnis			
		1	**X**	**2**	**Genauigkeit**
Prognose	**1**	**60**	32	26	50,85 %
	X	0	**0**	0	0,00 %
	2	9	6	**20**	57,15 %
	Trefferquote	86,96 %	0,00 %	43,48 %	**52,29 %**

Tabelle 8: Wahrheitsmatrix für Entscheidungsbäume

Eine Zeile in der Wahrheitsmatrix entspricht hierbei einer Vorhersage (d. h. das Ergebnis '1' wurde 118 mal vorhergesagt, das Ergebnis 'X' kein einziges mal, das Ergebnis '2' insgesamt 35 mal); mit den Spalten werden die tatsächlichen Ergebnisse abgebildet, so dass eine Zelle (i,j) aussagt, wie oft das Ergebnis i vorhergesagt wurde mit dem tatsächlichen Ergebnis j (vgl. Bramer, 2007, S. 89). In der Tabelle oben war von den 118 Fällen, in denen ein Heimsieg prognostiziert wurde, in 60 Fällen tatsächlich ein Heimsieg vorhanden, während in 32 Fällen ein Unentschieden und in 26 Fällen ein Sieg der Gastmannschaft vorhanden war.

Die Genauigkeit der Prognose eines Heimsieges liegt somit nur bei 50,85 %, da in 118 Fällen ein Heimsieg vorhergesagt wurde, wobei es nur in 60 Fällen tatsächlich einen Heimsieg gab. Die Trefferquote für Heimsiege liegt hingegen bei 86,96 %, da es insgesamt 69 Heimsiege gab, von denen 60 korrekt vorhergesagt wurden.

In der Diagonale kann somit abgelesen werden, wie oft korrekte Ergebnisse vorhergesagt wurden (vgl. Bramer, 2007, S. 89); im vorliegenden Fall wurde für 80 der 153 Spiele der korrekte Spielausgang prognostiziert, was zu einer Gesamt-Genauigkeit von 52,29 % führt.

Obwohl im Entscheidungsbaum Blätter mit dem Ergebnis Unentschieden existieren, wurden diese nie erreicht, d. h. das Ergebnis Unentschieden wurde nie vorhergesagt. Daher liegen sowohl Genauigkeit als auch Trefferquote für dieses Ergebnis bei 0 %.

In der Wahrheitsmatrix sagt die *Genauigkeit* somit aus, wie oft die Prognose eines bestimmten Ereignisses korrekt ist; während die Trefferquote aussagt, wie oft ein bestimmtes Ereignis korrekt vorhergesagt wird. Um auf die trivialen Prognosen aus Abschnitt 5.7.1 zurückzukommen: bei der Strategie, immer das häufigste Ereignis zu verwenden, wäre die Trefferquote für das Ergebnis Heimsieg bei 100 %, da jeder Heimsieg korrekt vorhergesagt wird, während die Trefferquote für Unentschieden und Siege der Gastmannschaften bei 0 % liegen würde, da diese nie vorhergesagt werden. Bei der Verwendung von Entscheidungsbäumen wird die Trefferquote für Heimsiege somit von 100 % auf etwa 87 % verschlechtert, während die Trefferquote für Siege der Gastmannschaft von 0 % auf etwa 43,5 % verbessert wurde; die Genauigkeit der gesamten Prognose stieg von 45,10 % auf 52,29 %.

5.7.3 Künstliche neuronale Netze

Beschreibung des Algorithmus

Mit künstlichen neuronalen Netzen wird versucht, den Aufbau des menschlichen Gehirns nachzubauen bzw. das menschliche Nervensystem zu simulieren, welches aus dem Zusammenschluss mehrerer Milliarden Nervenzellen (Neuronen) besteht (vgl. Aggarwal, 2015, S. 326). Diese Neuronen sind Aggarwal zufolge über sog. Synapsen miteinander verbunden, die sich im Rahmen eines Lern-Prozesses bilden bzw. ihre Stärke verändern; beim Aufbau eines künstlichen neuronalen Netzes wird nun versucht, diesen Prozess des Lernens zu simulieren. Der abstrakte Aufbau eines Neurons, wie es in künstlichen neuronalen Netz implementiert wird, befindet sich in Abbildung 10.

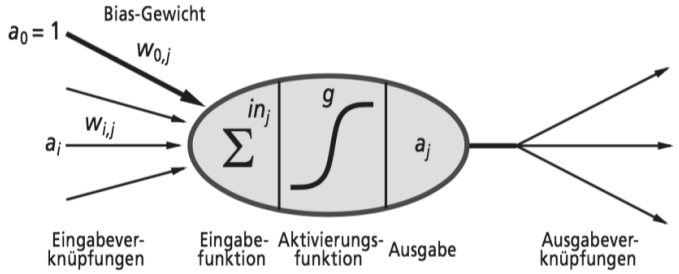


Abbildung 10: Aufbau eines Neutrons (Russell und Norvig, 2004, S. 846)

Im Folgenden die Erläuterung der einzelnen Funktionen eines Neutrons, angelehnt an die Ausführungen von Russell und Norvig (2004, S. 846 f.).

Die einzelnen Neuronen sind, wie bereits dargestellt, über Synapsen miteinander verbunden. Im Modell werden diese Verknüpfungen durch Pfeile (a) dargestellt, die für gerichtete Verknüpfungen stehen. Da die Synapsen in der Realität unterschiedliche Stärken besitzen, werden die einzelnen Verknüpfungen mit Gewichten (w) versehen.

Die Verknüpfungen der Neuronen ermöglichen es, sog. *Aktivierungen* zwischen den Neuronen zu transportieren. Die einzelnen Neuronen führen interne Berechnungen durch, um festzustellen, ob eine Aktivierung erfolgt.

Jedes Neuron nimmt nun im ersten Schritt die Eingaben der vorgelagerten Neuronen entgegen und berechnet im Rahmen der Eingabefunktion deren Summe:

$$\text{in}_j = \sum_{i=0}^{n} w_{i,j} a_i$$

Auf Basis dieser Summe erfolgt anschließend die Anwendung der *Aktivierungsfunktion*, um festzustellen, ob das Neuron aktiviert wird. Hierzu wird üblicherweise entweder eine Schwellenwertfunktion oder eine logistische Funktion angewandt.

$$a_j = g(\text{in}_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$$

Im Anschluss daran wird die Weitergabe – sofern eine Aktivierung erfolgt ist – des internen Aktivierungszustandes a an nachgelagerte Neuronen durchgeführt, mit welchen das Neuron über die Ausgabeverknüpfungen verbunden ist.

Auf eine detaillierte Darstellung der internen Vorgänge innerhalb der Neuronen wird verzichtet, im Folgenden wird nun noch dargestellt, wie einzelne Neuronen innerhalb eines neuronalen Netzes miteinander verbunden werden. Hierzu gibt es Russell und Norvig (2004, S. 847) zufolge zwei verschiedene Möglichkeiten: einerseits ist ein sog. **vorwärtsgerichtetes Netz** möglich, in welchen die möglichen Verbindungen zwischen Neuronen beschränkt sind, wodurch verhindert wird, dass Zyklen entstehen; im Gegensatz dazu können Neuronen in einem **rückgekoppelten Netz** beliebige Verbindungen zu anderen Neuronen und zu sich selbst besitzen. Der Vorteil des zweiten Ansatzes liegt nach Russell und Norvig darin, dass diese das menschliche Gehirn besser simulieren können, da durch die Ermöglichung von Zyklen ein lokales Kurzzeitgedächtnis unterstützt wird. Aufgrund dessen, dass üblicherweise vorwärtsgerichtete Netze verwendet werden (das verwendete Framework *RapidMiner* implementiert auch nur diesen Ansatz, vgl. hierzu RapidMiner, 2016a), wird die Darstellung im Folgenden auf diesen Ansatz beschränkt.

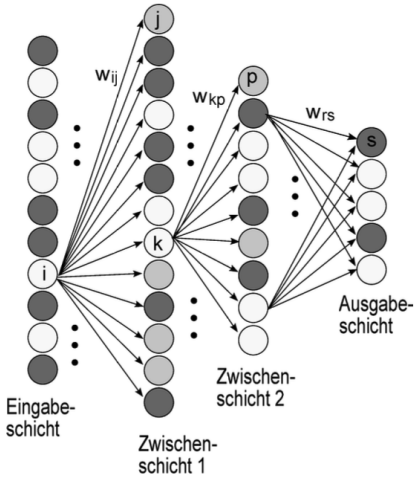


Abbildung 11: Beispiel für den Aufbau eines vorwärtsgerichteten Netzes (Cleve und Lämmel, 2014, S. 120)

In Abbildung 11 wird ein Beispiel für ein vorwärtsgerichtetes Netz dargestellt. Grundsätzlich ist dieses Cleve und Lämmel (2014, S. 120) zufolge in Schichten angeordnet, beginnend mit der Eingabeschicht, mit der die verschiedenen Attribute eines Datensatzes eingespeist werden können; endend mit der Ausgabeschicht, mit welchem das Ergebnis für den eingespeisten Datensatzes ausgegeben wird. Zwischen diesen beiden Schichten gibt es mindestens eine Zwischenschicht; in diesen sind Neuronen immer nur mit Neuronen von vor- oder nachgelagerten Schichten verbunden (vgl. Cleve und Lämmel, 2014, S. 120 f.). Um ein neuronales Netz aufzubauen, wird Cleve und Lämmel zufolge im vorliegenden Fall des überwachtes Lernen üblicherweise der sog. *Backpropagation-of-error*-Algorithmus angewandt, bei welchem nacheinander Datensätze der Trainingsdatenmenge eingespeist

werden, anschließend wird festgestellt, ob eine Abweichung des Ergebnisses durch das Netz zu der bekannten Klassifikation vorhanden ist. Für den Fall, dass ein Fehler vorhanden ist, wird das neuronale Netz angepasst, indem neue Verknüpfungen hinzugefügt oder die Gewichtungen bestehender Verknüpfungen angepasst werden, oder indem die Aktivierungsfunktionen einzelner Neuronen verändert werden, mit dem Ziel, dass der Fehler beim erneuten Einspeisen des Datensatzes verringert wird.

Für ein grundsätzliches Verständnis sollten die dargestellten Informationen genügen, für eine detailliertere Darstellung sei an dieser Stelle auf die zitierte Literatur verwiesen.

Analyseergebnisse

Die Gesamt-Genauigkeit beim Einsatz künstlicher neuronaler Netze beträgt 51,63 %, in Tabelle 9 wird die Wahrheitsmatrix abgebildet. Der Algorithmus verwendet hierbei die beiden Attribute Marktwertdifferenz und Gegentordifferenz; die entstandenen Modelle befinden sich in Anhang A.2.

Wie aus der Wahrheitsmatrix ablesbar ist, wird wie beim Einsatz von Entscheidungsbäumen ebenfalls nie ein Unentschieden vorhergesagt, obgleich der Algorithmus intern diesen Spielausgang berücksichtigt (siehe Modell des neuronalen Netzes im Anhang).

Im Gegensatz zu den Ergebnissen des Einsatzes von Entscheidungsbäumen ist die Gesamt-Genauigkeit des Modells schlechter, die Genauigkeit der Vorhersage von Auswärtssiegen ist jedoch besser, da in etwa 73 % der vorhergesagten Fälle tatsächlich ein Auswärtssieg vorkam (bei Entscheidungsbäumen: 57 %).

		tatsächliches Ergebnis			
		1	**X**	**2**	**Genauigkeit**
Prognose	**1**	**68**	35	35	49,28 %
	X	0	**0**	0	0,00 %
	2	1	3	**11**	73,33 %
	Trefferquote	98,55 %	0,00 %	23,91 %	**51,63 %**

Tabelle 9: Wahrheitsmatrix für künstliche neuronale Netze

Auf eine tiefergehende Erläuterung der Wahrheitsmatrizen wird an dieser Stelle und im Folgenden verzichtet, da die Aussagen der dargestellten Inhalte bereits im vorherigen Abschnitt genügend erläutert wurden.

5.7.4 Naive Bayes

Beschreibung des Algorithmus

Der Naive-Bayes-Algorithmus basiert auf dem Satz von Bayes und gehört zu den probabilistischen Bayes-Klassifikatoren (vgl. Runkler, 2015, S. 93). Mit Hilfe des Satzes von Bayes kann man die bedingte Wahrscheinlichkeit eines Ereignisses errechnen, weshalb der Algorithmus die Wahrscheinlichkeiten aller Ereignisse berechnet und das Ereignis auswählt, das die größte Wahrscheinlichkeit besitzt (vgl. Aggarwal, 2015, S. 306).

Der **Satz von Bayes** bezieht sich auf die *bedingte* Wahrscheinlichkeit $P(A \mid B)$ des Eintretens von A unter der Voraussetzung, dass B eingetreten ist (vgl. Tan u. a., 2006, S. 230):

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Während in der dargestellten Formel nur die Abhängigkeit von einem Attribut (B) dargestellt wird, lässt sich diese Formel so anpassen, dass beliebig viele Attribute als Bedingungen eingefügt werden können.

Der Naive-Bayes-Algorithmus basiert auf diesem Theorem und verwendet dies, indem die Wahrscheinlichkeiten von jedem der verwendeten Attribute für die möglichen Klassifikationen errechnet werden (vgl. Tan u. a., 2006, S. 230 f.). Auf eine tiefergehende Erläuterung wird an dieser Stelle verzichtet, da dies den Rahmen sprengen würde.

Der Algorithmus wird als „naiv“ bezeichnet, da hierbei davon ausgegangen wird, dass die Attribute untereinander keine Abhängigkeit besitzen, sondern nur vom Klassenattribut (hier: Spielergebnis) abhängig sind (vgl. Aggarwal, 2015, S. 310). Dies ist in der Realität nicht gegeben, ein einfaches Beispiel im Bereich des Fußballs ist der Tabellenplatz und die erreichte Punktzahl, da der Tabellenplatz sich mit der Anzahl der erreichten Punkte verbessert. Der Naive-Bayes-Algorithmus errechnet daher die einzelnen bedingten Wahrscheinlichkeiten, ohne dass eine eventuell vorhandene Korrelation zwischen den Attributen betrachtet wird.

Im Folgenden ein einfaches Beispiel unter Verwendung des Attributes Tabellenplatz der Heimmannschaft auf Basis der in dieser Arbeit verwendeten Datenbasis. Dies dient nur zur Veranschaulichung, in Wirklichkeit wäre die Qualität der Prognosen vermutlich recht schlecht, da nur die Heimmannschaft betrachtet wird. Des Weiteren wird der Einfachheit halber nur ein einzelnes Attribut betrachtet, bei der wirklichen Anwendung werden mehrere Attribute kombiniert.

Die Wahrscheinlichkeit für das Ergebnis Heimsieg einer Mannschaft, die aktuell auf dem ersten Platz steht, ist durch die Anwendung des Bayes-Theorem errechenbar. Gesucht ist daher die Wahrscheinlichkeit für einen Heimsieg, unter der Voraussetzung, dass der Tabellenplatz der Heimmannschaft 1 ist. Im Folgenden die Formel hierfür (analog für

Unentschieden oder Niederlagen, siehe weiter unten).

$$P(\text{Erg} = 1 \mid \text{Tbl} = 1) = \frac{P(\text{Tbl} = 1 \mid \text{Erg} = 1) * P(\text{Erg} = 1)}{P(\text{Tbl} = 1)}$$

Im ersten Schritt müssen nun die Wahrscheinlichkeiten der drei möglichen Spielausgänge berechnet werden, wobei hier alle Daten der Trainingsdatenmenge betrachtet werden, da das Modell anhand dieser erzeugt wird.

$$\begin{aligned} P(\text{Erg} = 1) &= \frac{758}{1645} \approx 0.461 \\ P(\text{Erg} = X) &= \frac{386}{1645} \approx 0.235 \\ P(\text{Erg} = 2) &= \frac{501}{1645} \approx 0.305 \end{aligned} \tag{1}$$

Anschließend wird die Wahrscheinlichkeit, dass der Tabellenplatz der Heimmannschaft den Wert 1 besitzt, errechnet. Insgesamt stand 99mal die Heimmannschaft auf Platz 1:

$$P(\text{Tbl} = 1) = \frac{99}{1645} \approx 0.060$$

Nun muss die bedingte Wahrscheinlichkeit für die Tabellenposition errechnet werden, wenn ein bestimmtes Ergebnis vorhanden ist. Hierzu wird die Anzahl der Spiele, bei denen die Heimmannschaft auf Platz 1 stand (79) und das Spiel gewonnen hat, mit der Anzahl aller Heimsiege (758) verrechnet; analog hierzu wird dies für Unentschieden und Niederlagen durchgeführt:

$$\begin{aligned} P(\text{Tbl} = 1 \mid \text{Erg} = 1) &= \frac{79}{758} \approx 0.104 \\ P(\text{Tbl} = 1 \mid \text{Erg} = X) &= \frac{10}{386} \approx 0.026 \\ P(\text{Tbl} = 1 \mid \text{Erg} = 2) &= \frac{10}{501} \approx 0.020 \end{aligned}$$

Nun kann die oben dargestellte Formel angewandt werden:

$$\begin{aligned} P(\text{Erg} = 1 \mid \text{Tbl} = 1) &= \frac{P(\text{Tbl} = 1 \mid \text{Erg} = 1) * P(\text{Erg} = 1)}{P(\text{Tbl} = 1)} &&= \frac{0.104 \cdot 0.461}{0.060} \approx 0,799 \\ P(\text{Erg} = X \mid \text{Tbl} = 1) &= \frac{P(\text{Tbl} = 1 \mid \text{Erg} = X) * P(\text{Erg} = X)}{P(\text{Tbl} = 1)} &&= \frac{0.026 \cdot 0.235}{0.060} \approx 0,102 \\ P(\text{Erg} = 2 \mid \text{Tbl} = 1) &= \frac{P(\text{Tbl} = 1 \mid \text{Erg} = 2) * P(\text{Erg} = 2)}{P(\text{Tbl} = 1)} &&= \frac{0.020 \cdot 0.305}{0.060} \approx 0,102 \end{aligned}$$

Somit liegt die Wahrscheinlichkeit eines Heimsieges bei etwa 79,9 %, während die Wahrscheinlichkeiten für Unentschieden und Niederlage bei jeweils etwa 10,2 % liegen (Anmerkung: Aufgrund von Rundungsfehlern ergibt die Summe mehr als 100 %). Der Naive-Bayes-Algorithmus würde nun das Ergebnis mit der größten Wahrscheinlichkeit auswählen, d. h. im vorliegenden Beispiel würde er immer einen Heimsieg prognostizieren, was

in Anbetracht dessen, dass die Heimmannschaft auf Platz 1 der Tabelle steht, in vielen Fällen eine realistische Prognose ist.

Analog zur obigen Berechnung müssten die Wahrscheinlichkeiten für weitere Tabellenpositionen errechnet werden, da mit der oben errechneten Regel nur Spiele vorhergesagt werden können, bei denen die Heimmannschaft auf Platz 1 steht; das Ziel wäre jedoch ein ganzheitliches Modell, um alle möglichen Spiele betrachten zu können.

Analyseergebnisse

Im Gegensatz zum Beispiel oben, wo für das Attribut Tabellenplatz eine beschränkte Anzahl an möglichen Werten vorhanden ist (1–18, nur ganzzahlig), sind andere verwendete Attribute rational (z. B. durchschnittliche Punkte), womit es viele mögliche und insbesondere evtl. vorab unbekannte Werte gibt, weshalb der oben dargestellte Ansatz in einer abgewandelten Form angewandt wird. Deshpande (2012) zufolge wird im RapidMiner intern die Gaußsche Normalverteilung zur Berechnung der einzelnen Wahrscheinlichkeiten verwendet.

Im Rahmen der Optimierung hat sich herausgestellt, dass der naive Bayes-Algorithmus unter Verwendung der beiden Attribute 'Differenz der Marktwerte' und 'Differenz der Gegentore' gute Ergebnisse bringt; im Anhang A.1 befinden sich die durch den RapidMiner verwendeten Gauß-Kurven für die beiden verwendeten Attribute.

In Tabelle 10 wird die Wahrheitsmatrix der Ergebnisse des Einsatzes der naiven Bayes-Klassifikators. Die Gesamt-Genauigkeit liegt hier bei 54,90 %, da für 84 der 153 Spiele ein korrekter Spielausgang vorhergesagt wurde.

		tatsächliches Ergebnis			
		1	**X**	**2**	**Genauigkeit**
Prognose	**1**	**63**	29	25	53,85 %
	X	0	**0**	0	0,00 %
	2	6	9	**21**	58,33 %
	Trefferquote	91,30 %	0,00 %	45,65 %	**54,90 %**

Tabelle 10: Wahrheitsmatrix für Naive Bayes

Im Vergleich zum Einsatz künstlicher neuronaler Netze ist die Genauigkeit der Prognose von Auswärtssiegen schlechter, im Gegenzug ist die Genauigkeit von Heimsiegen etwas besser; insgesamt führt dies zu einer deutlichen Verbesserung der Genauigkeit des Modells, da die Klasse 'Heimsieg' wesentlich mehr Datensätze beinhaltet.

5.7.5 Logistische Regression

Beschreibung des Algorithmus

Bei der Regression wird versucht, den Zusammenhang zwischen einer Zielvariable und einem oder mehreren Attributen von Datensätzen zu beschreiben. Die einfachste Form der Regression ist hierbei die sog. univariate lineare Regression, bei welcher Eingabe und Ausgabe aus jeweils einer Variable bestehen (vgl. Russell und Norvig, 2004, S. 836). Das Ergebnis ist hierbei folgende Funktion, die eine Gerade darstellt (vgl. Runkler, 2015, S. 67 f.):

$$y = b + a * x$$

Hierbei stellt x den Wert der Eingabe, a die Gewichtung dieser Eingabe und b eine Konstante, den Ordinatenabschnitt oder Intercept dar; y ist dar geschätzte Wert der Zielvariablen; die beiden Parameter a und b werden hierbei während der Lernphase geschätzt (vgl. Runkler, 2015, S. 68).

In einer erweiterten Form, der sog. multivariaten linearen Regression wird ein Datensatz nicht durch ein einzelnes sondern durch mehrere Attribut beschrieben, so dass ein „n-elementiger Vektor“ vorhanden ist (vgl. Russell und Norvig, 2004, S. 836). Die Formel der Geraden lautet Runkler (vgl. 2015, S. 69 f.) zufolge daher folgendermaßen:

$$y = b + a_1 * x_1 + a_2 * x_2 + ... + a_n * x_n = b + \sum_i a_i * x_i$$

Ein einfaches (fiktives) Beispiel: grundsätzlich wird davon ausgegangen, dass der Marktwert eines Teams mit dem Erfolg, d. h. mit der Anzahl erreichten Punkte in einer Saison korreliert, da die teuersten Teams üblicherweise um die Meisterschaft spielen, während Teams mit geringeren Marktwerten häufig um den Abstieg spielen. Eine mögliche Funktion, welche die Abhängigkeit der erreichten Punkte vom Marktwert eines Teams (in Mio.) abbildet, könnte daher folgendermaßen lauten:

$$y = 30 + 0.2 * x_{\text{Marktwert}}$$

Ein Team mit dem Marktwert 10 Millionen Euro würde in diesem Modell voraussichtlich 32 Punkte erreichen, ein Team mit 100 Millionen Euro 50 Punkte und ein Team mit 200 Millionen Euro 70 Punkte.

An diesem Beispiel wird auch schon ein Problem bei dem Einsatz der linearen Regression zur Prognose ersichtlich: der Output eines linearen Regressionsmodells besitzt einen kontinuierlichen Wertebereich, im vorliegenden Fall der Prognose von Spielergebnissen muss jedoch eine diskrete Variable vorhergesagt werden, da es nur drei mögliche Spielergebnisse gibt. Um Regressionsmodelle auch zur Prognose diskreter Werte verwenden zu können, können Schwellenwertfunktionen eingebunden werden, so dass der kontinuierliche Wertebereich in zwei Bereiche unterteilt wird, denen jeweils ein Wert einer binomialen Variablen zugewiesen werden kann (vgl. Russell und Norvig, 2004, S. 841 f.). Hierbei befindet

sich in Abbildung 12 auf der linken Seite eine exemplarische Schwellenwertfunktion, die negative Werten den Output 0 zuweist, positiven Werten hingegen den Wert 1.

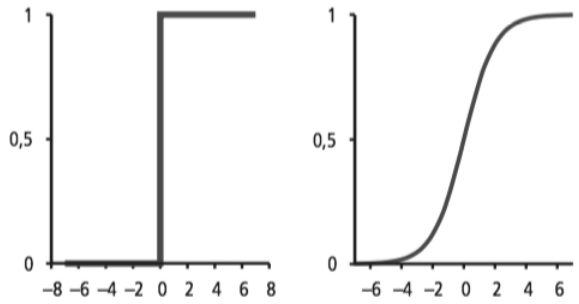


Abbildung 12: links: Schwellenwertfunktion; rechts: logistische Funktion (Russell und Norvig, 2004, S. 844)

Der Verlauf dieses Schwellenwerts verursacht jedoch diverse Probleme (vgl. hierzu Russell und Norvig, 2004, S: 843 f.); ein für unsere Prognose zu berücksichtigendes Problem ist hierbei, dass der Schwellenwert eine fixe Zahl ist, anhand der eine Zuweisung zu den zwei Outputs stattfindet. Im Beispiel oben würde daher ein Wert von –0.001 dem Output 0 zugewiesen werden, während +0.001 zum Output 1 führen würde. Oft wird jedoch ein feiner abgestufter Verlauf dieser Zuweisung benötigt, der auch für nah an dieser Schranke liegende Fälle sinnvolle Ergebnisse liefert (vgl. Russell und Norvig, 2004, S. 844). Daher wird oftmals eine *logistische* Funktion angewandt werden, ein Beispiel hierfür ist die rechte Funktion in der Abbildung oben. Diese Funktion liefert nun beliebige Werte im Bereich zwischen 0 und 1 zurück, was als Wahrscheinlichkeitsverteilung für die beiden möglichen Werte verwendet werden kann (vgl. Russell und Norvig, 2004, S. 844). Auf das Beispiel oben angewandt würden nun beide Werte in der Nähe von 0 etwa einen Output von 0.5 liefern, d. h. eine Wahrscheinlichkeit von je 50 % für beide Ausgänge anstatt der fixen Zuweisung zu einem der beiden Outputs.

Auf eine weitere Detaillierung wird an dieser Stelle verzichtet, zum Verständnis sollten die oben dargestellten Funktionen reichen; weitere Details würden zu sehr in die theoretische Statistik eintauchen; daher sei an dieser Stelle auf die zitierte Literatur verwiesen

Mit den oben dargestellten Möglichkeiten können nur binomiale Variablen prognostiziert werden, oft sind jedoch multinomiale Variablen vorhanden; im Fall der Prognose von Fußballspielen müssen beispielsweise drei mögliche Ergebnisse berücksichtigt werden. Eine einfache Möglichkeit, dieses Problem zu lösen, wäre, nur zwei mögliche Ergebnisse zu berücksichtigen und beispielsweise Unentschieden komplett zu ignorieren, was aber recht wenig sinnvoll wäre, da derartige Modelle die Realität nicht mehr korrekt wiedergeben würden. Es besteht jedoch auch die Möglichkeit, das Problem in mehrere binomiale Teilprobleme umzuwandeln, so dass die Algorithmen wie gewohnt angewandt werden können (vgl. Tan u. a., 2006, S. 306). Hierbei gibt es Tan u. a. zufolge zwei verschiedene Ansätze: den sog. *One-against-Rest*-Ansatz, bei dem für jede mögliche Klassifikation ein binomia-

les Problem erzeugt wird, in welchem die betrachtete Klassifikation zum Wert 1 führt, während alle anderen Klassifikationen mit dem Wert 0 belegt werden. Die zweite Möglichkeit ist der *One-against-One*-Ansatz, bei dem jeweils zwei Klassifikationen betrachtet werden, wobei eine dem Wert 0 und die andere mit 1 belegt wird, während die restlichen Klassifikationen ignoriert werden.

Angewandt auf unser Fußball-Problem würden beim One-against-Rest-Ansatz die Modelle für '1' gegen den Rest, 'X' gegen den Rest und '2' gegen den Rest erzeugt werden, während bei One-against-One die Modelle '1' gegen den 'X', 'X' gegen den '2' und '1' gegen den '2' erzeugt werden. In Tests auf dem Trainingsdatenbestand hat sich der zweite Ansatz besser bewährt, weshalb dieser im Rahmen dieser Arbeit angewandt wird.

Um eine Prognose erstellen zu können, müssen die entstandenen Modelle logischerweise auch wieder zusammengeführt werden, da für jeden Datensatz eine einheitliche Prognose erzeugt werden muss. Hierzu werden alle in Frage kommenden Modelle auf einen Datensatz angewandt, um danach eine Abstimmung (*voting*) durchzuführen, bei welcher diejenige Klassifikation „gewinnt", welche die meisten Stimmen erhält; alternativ können die Stimmen jedoch auch zur Berechnung einer Wahrscheinlichkeit für die verschiedenen Ausgänge verwendet werden (vgl. Tan u. a., 2006, S. 307).

Analyseergebnisse

Die Genauigkeit des entstandenen Modells liegt bei 51,63 %, die Wahrheitsmatrix hierzu befindet sich in Abbildung 11. Der Algorithmus verwendet hierbei drei Attribute: Marktwertdifferenz, Gegentordifferenz und die Differenz der Formtabellenpositionen der beiden Mannschaften. In Anhang A.3 befinden sich die detaillierten Ergebnisse in Form der Gewichte der verwendeten Attribute für die Regressionsfunktionen.

Im Vergleich zu den bisherigen Modellen werden Auswärtssiegen hierbei zwar seltener prognostiziert, jedoch ist die Genauigkeit der Prognose von Auswärtssiegen besser, da diese von den 13 Fällen nur zweimal falsch lag.

		tatsächliches Ergebnis			
		1	**X**	**2**	**Genauigkeit**
Prognose	**1**	**68**	37	35	48,57 %
	X	0	**0**	0	0,00 %
	2	1	1	**11**	84,62 %
	Trefferquote	98,55 %	0,00 %	23,91 %	**51,63 %**

Tabelle 11: Wahrheitsmatrix für die logistische Regression

5.8 Zusammenfassung

Algorithmus	**Genauigkeit**	**verwendete Attribute**
Zufällige Prognose	33,33 %	keine Attribute
Immer häufigstes Ergebnis	45,10 %	keine Attribute
Entscheidungsbaum	52,29 %	- Differenz Marktwert - Tabellenplatz Heim - Punkteschnitt Form Heim
Künstliche neuronale Netze	51,63 %	- Diff. Marktwert - Diff. Gegentore
Naive Bayes	54,90 %	- Diff. Marktwert - Diff. Gegentore
Logistische Regression	51,63 %	- Diff. Marktwert - Diff. Gegentore - Diff. Tabellenplatz Form

Tabelle 12: Ergebnisse der einzelnen Algorithmen

In der Tabelle 12 werden die Prognoseergebnisse abschließend aufgelistet. Insgesamt wurden 15 Attribute erarbeitet, von denen die Algorithmen dann jeweils im Rahmen eines Optimierungsprozesses bis zu drei Attribute auswählen konnten. Alles in allem wurden hierbei jedoch nur fünf dieser Attribute ausgewählt:

- Differenz der Marktwerte der beiden Mannschaften,
- Differenz der Gegentore der beiden Mannschaften,
- Tabellenposition der Heimmannschaft,
- Punkteschnitt der Heimmannschaft in der Formtabelle,
- Differenz der Tabellenpositionen der beiden Mannschaften.

Von jedem Algorithmus wurde dabei die Differenz der Marktwerte verwendet, dreimal wurde die Differenz der Gegentore ausgewählt, während die anderen drei Attribute jeweils nur einmal ausgewählt wurde. Zwei dieser Attribute beziehen sich auf die Heimmannschaft, während sich drei auf die Differenz zwischen Heimmannschaft und Gastmannschaft beziehen; dies deutet an, dass es sinnvoll war, auch die Differenzen der Attribute heranzuziehen.

Eine weitere Auffälligkeit ist, dass keiner der Algorithmen Spiele mit dem Ergebnis Unentschieden korrekt prognostiziert hat; ganz im Gegenteil, die Algorithmen haben sogar erst gar nicht versucht, dieses Ergebnis vorherzusagen. Ein Grund hierfür ist sicherlich, dass das Spielergebnis Unentschieden mit einer Wahrscheinlichkeit von etwa 25 % das seltenste Ergebnis ist, jedoch sind Auswärtssiege mit etwa 30 % auch nicht deutlich häufiger

anzutreffen. In anderen wissenschaftlichen Arbeiten ist dieses Problem jedoch auch anzutreffen, weshalb teilweise auch die Modellierung einer Fußball-Prognose als 2-Klassen-Problem durchgeführt wird, so dass im Vorhinein bereits das Ergebnis Unentschieden ignoriert wird.

Die Ergebnisse liegen - lässt man die zufällige Prognose außer Acht - zwischen 45,10 % und 54,90 %. Die Strategie, immer das häufigste Ergebnis zu wählen, diente dazu, die anderen Prognoseergebnisse besser einschätzen zu können. Im Vergleich mit dieser wurde somit eine Verbesserung um knapp 10 % erzielt. Insgesamt sind die Prognoseergebnisse mit einer Genauigkeit von etwa 55 % durchschnittlich, da in vergleichbaren Arbeiten üblicherweise Genauigkeiten von 50 % bis etwa 60 % erreicht werden.

6 Fazit

Zum Abschluss wird in diesem Kapitel die gesamte Arbeit kurz zusammengefasst, außerdem wird versucht, ein Fazit zu ziehen und einen Ausblick auf mögliche Erweiterungen der durchgeführten Analysen werfen.

In dieser Arbeit wurde dargestellt, in welchen Bereichen im Fußball Datenanalysen bereits eingesetzt werden und welche Vorteile dadurch ermöglicht werden. Außerdem wurde durch die Anwendung verschiedener Data-Mining-Verfahren auf Daten der deutschen Bundesliga dargestellt, inwiefern sich die Ergebnisse von Fußballspielen prognostizieren lassen. Die Vorgehensweise hierbei hat sich an den *Knowledge Discovery in Databases*-Prozess angelehnt: beginnend mit der Datenerfassung und -Vorverarbeitung, hin zur Durchführung von insgesamt vier Data-Mining-Algorithmen auf diesem Datenbestand, abgeschlossen mit einer Evaluation der Ergebnisse.

Um auf die Forschungsfrage aus der Einleitung zurückzukommen: hat der Zufall einen zu großen Einfluss auf den Fußball, als dass es sinnvolle Möglichkeiten zum Einsatz gibt, oder lassen sich im Fußball sinnvolle Einsatzmöglichkeiten finden?

Diese Frage lässt sich abschließend nicht eindeutig beantworten. Im Fußball bieten sich mit Sicherheit viele Einsatzmöglichkeiten, die sinnvoll sind und die sich bereits innerhalb der Vereine bewährt haben. Hierzu zählt insbesondere der Ansatz der *Performance Analysis*, mit welchem die Leistungen und Entwicklungspotentiale von Spielern beurteilt werden können; außerdem ist der Einsatz von Analysen zur Kaderplanung üblich, d.h. zur Auswahl von Spielern, die verpflichtet oder abgegeben werden sollen.

Jedoch gibt es genauso Bereiche, in denen es nicht wirklich sinnvoll ist, auf Data-Mining-Ansätze zu setzen. So wurde in den durchgeführten Analysen versucht, Spielergebnisse vorherzusagen, wobei Genauigkeiten von bis zu 55% erreicht wurden, was eine deutliche Steigerung im Vergleich zu zufälligen Prognosen (33%) und zu trivialen Prognosen (45%) ist. Im Vergleich zu anderen wissenschaftlichen Arbeiten ist dieser Wert nur durchschnittlich, teilweise werden dort Genauigkeiten von bis zu 65% erreicht. Die Frage ist jedoch, inwiefern diese Ergebnisse sinnvoll verwendet werden können. Außerhalb der Vereine mag es Sinn machen, Spielergebnisse vorherzusagen, insbesondere im Bereich der Sportwetten, da sich die Wettanbieter auf Datenanalysen zur Berechnung von Gewinnwahrscheinlichkeiten fokussieren, anhand welcher die Wettquoten festgelegt werden.

Für den Einsatz innerhalb von Vereinen eignen sich derartige Prognosen vermutlich weniger, da diese Genauigkeiten zu schlecht sind, als dass sie sich für betriebswirtschaftliche Einsatzmöglichkeiten qualifizieren. Das Problem liegt darin, dass der Zufall definitiv einen großen Einfluss besitzt; so kommt es regelmäßig vor, dass der Außenseiter ein Spiel gewinnt, oder dass Mannschaften absteigen, die in den Jahren zuvor stets gute Leistungen erbracht haben. Mit dem Zufall alleine lässt sich dies auch nicht begründen, viel mehr gibt es verschiedene Einflüsse, die nicht messbar sind, wie z.B. psychologische Faktoren oder Verletzungen von Spielern.

Aus einer wissenschaftlichen Sichtweise heraus wäre es natürlich erstrebenswert, Spiele möglichst genau vorhersagen zu können, von den Fans wird der Einsatz derartiger Analysen jedoch auch kritisiert, da die Spannung im Fußball eben genau darin liegt, dass Spiele (zum Teil) unvorhersehbar sind und auch der Außenseiter gewinnen kann.

Ausblick

Wie dargestellt wurde eine Einschränkung der Algorithmen auf die Verwendung von maximal drei Attributen vorgenommen, was logischerweise den Nachteil hat, dass die entstandenen Modelle recht einfach sind, jedoch wiegt der Vorteil der Begrenzung von Overfitting dies auf, da die Ergebnisse hierdurch verbessert werden konnten.

Interessant wäre nun, inwiefern dieses Problem anderweitig gelöst werden kann, so dass die Algorithmen beliebig viele Attribute verwenden können, ohne Probleme im Bezug auf Overfitting zu besitzen. Innerhalb der Analysen dieser Arbeit wurde keine bessere Lösung gefunden, als die Einschränkung hinsichtlich der Anzahl der Attribute, jedoch gibt es beispielsweise auch die Möglichkeit der Verwendung von Algorithmen, die große Anzahlen von Attributen besser verarbeiten können. Hier wurde dieser Ansatz jedoch nicht weiter beachtet, da das verwendete Tool RapidMiner hierfür nur einen Algorithmus bereitstellt (SVM), welcher bei der Anwendung jedoch keine Verbesserung der Ergebnisse brachte.

Darüber hinaus wäre es interessant, weitere Attribute als Grundlage für die Prognosen zu verwenden, beispielsweise Kennzahlen mit einem Bezug zur Leistung einzelner Spieler und deren Wichtigkeit für die Mannschaft. Dies würde dazu führen, dass Prognosen während laufenden Spielen ermöglicht werden, da es Dienste gibt, die Live-Daten bereitstellen (vgl. Abschnitt 4.3). Die Umsetzung einer Prognose auf Spieler-Basis ist jedoch nicht mit frei verfügbaren Datenquellen umsetzbar; die kommerziellen Anbieter verlangen jedoch mehrere Tausend Euro für Datenzugriffe. Außerdem besitzt eine derartige Analyse große sportwissenschaftliche Anteile, da definiert werden muss, wie die Leistung von Spielern und deren Einfluss auf eine Mannschaften bewertet werden kann. Aus diesen Gründen hat sich dieser Teilbereich nicht für den Einsatz in dieser Arbeit geeignet, nichtsdestotrotz lägen hier potentielle Erweiterungsmöglichkeiten für den Einsatz von Prognose-Techniken im Fußball.

A Anhang

A.1 Ergebnisse Naive Bayes

In Abbildung 13 wird die Gaußkurve für die Marktwertdifferenz abgebildet. Wie bereits erläutert ist die Marktwertdifferenz positiv, falls die Heimmannschaft einen höheren Marktwert als die Gastmannschaft hat, und negativ, falls die Gastmannschaft teurer ist. Die Kurve für die Wahrscheinlichkeit eines Heimsieges ist hierbei rechts, was bedeutet, dass die Wahrscheinlichkeit für einen Heimsieg größer wird, je höher die Marktwertdifferenz ist. Die Kurve für Auswärtssiege ist leicht nach links verschoben, was bedeutet, dass die Wahrscheinlichkeit für einen Auswärtssieg größer wird, wenn die Marktwertdifferenz kleiner wird.

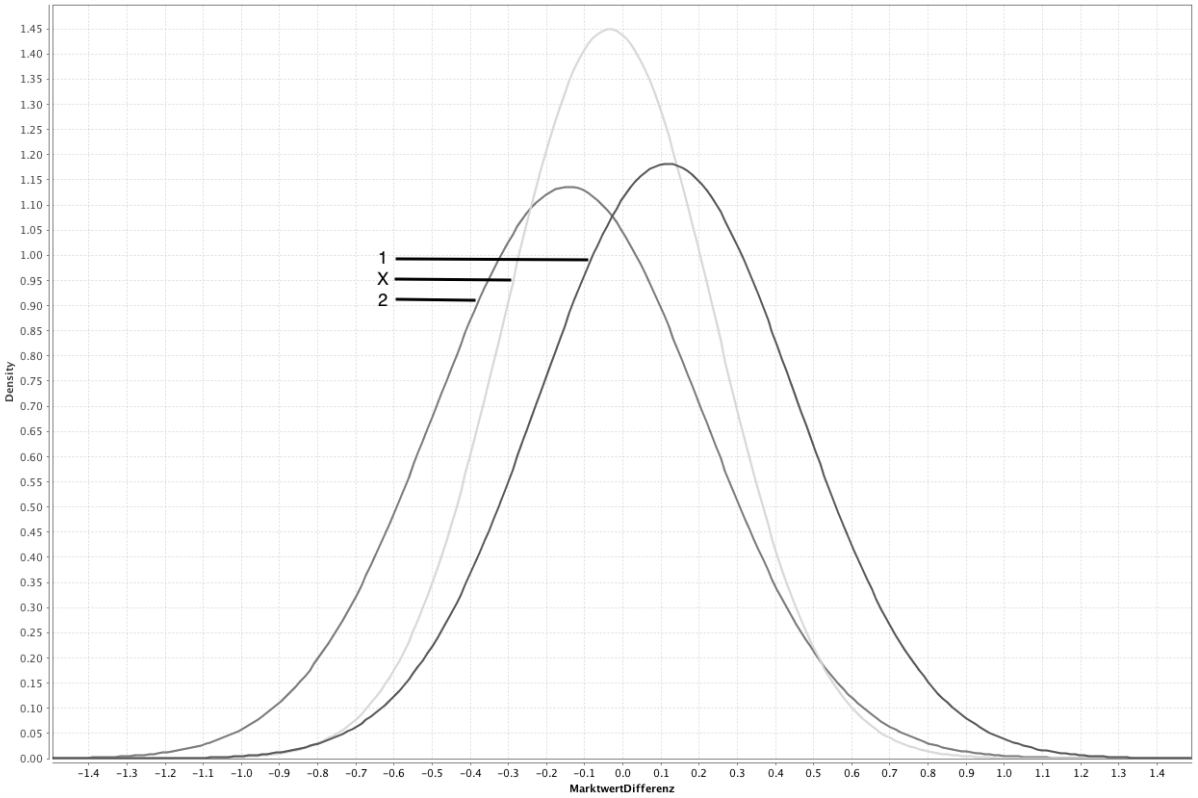


Abbildung 13: Gaußkurve des verwendeten Attributs Marktwertdifferenz

In Abbildung 14 befindet sich zudem die Gaußkurve des zweiten verwendeten Attributes Gegentordifferenz. Die Reihenfolge der Kurven ist dort im Vergleich zu der Reihenfolge der Marktwertdifferenz gedreht, da die Bedeutung einer Veränderung des Attributes gegensätzlich ist. Da eine geringere Gegentoranzahl für ein Team besser ist, hat ein negativer Wert der Gegentordifferenz beider Teams eine positive Bedeutung für das Heimteam, da das bedeutet, dass das gegnerische Team im Durchschnitt mehr Gegentore erhält.

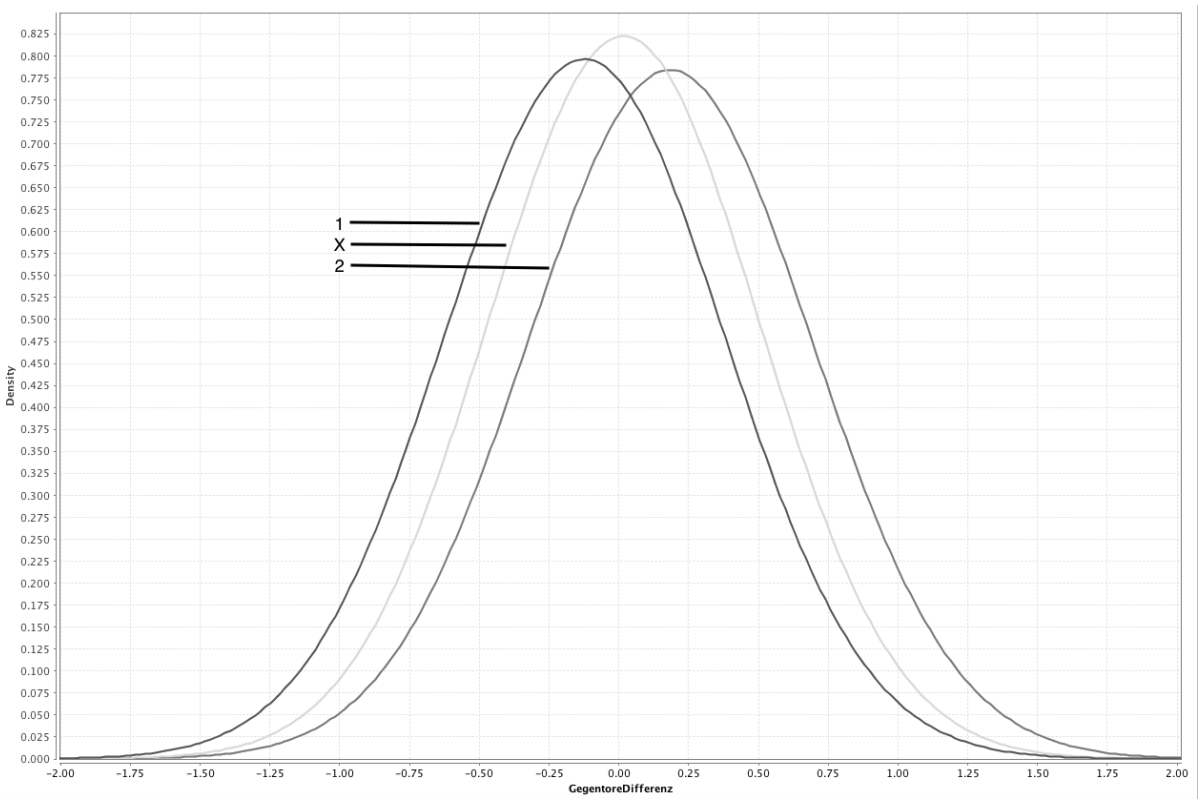


Abbildung 14: Gaußkurve des verwendeten Attributs Gegentordifferenz

Die Parameter der abgebildeten Gaußkurven befinden sich in Tabelle 13:

		Mittelwert	Standardabweichung
MarktwertDifferenz	**1**	0.116	0.338
	X	–0.035	0.275
	2	–0.143	0.351
GegentoreDifferenz	**1**	–0.123	0.501
	X	0.017	0.485
	2	0.183	0.509

Tabelle 13: Parameter der Gaußkurven

A.2 Ergebnisse des künstlichen neuronalen Netzes

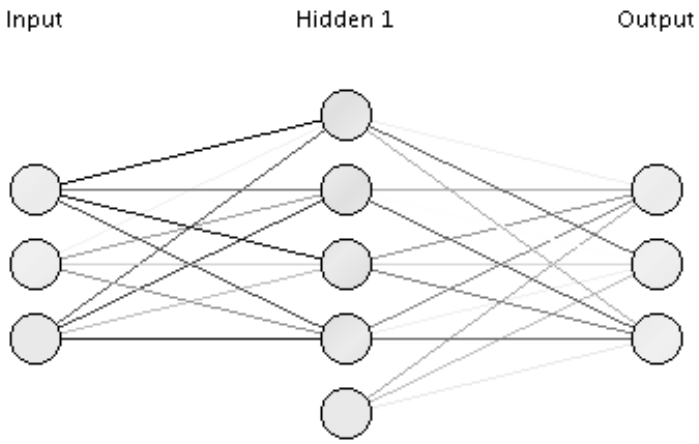


Abbildung 15: Ergebnismodell des künstlichen neuronalen Netzes

Anmerkung: die Stärken der Verbindungen der Neutronen werden durch die Graustufen in der Abbildung dargestellt.

ImprovedNeuralNet

```
Hidden 1                        Output
========                        ======

Node 1 (Sigmoid)                Class '1' (Sigmoid)
----------------                -------------------
MarktwertDifferenz: 4.542       Node 1: 0.245
GegentoreDifferenz: -0.206      Node 2: 1.401
Bias: -3.012                    Node 3: 1.874
                                Node 4: 2.116
Node 2 (Sigmoid)                Threshold: -1.516
----------------
MarktwertDifferenz: 2.373       Class 'X' (Sigmoid)
GegentoreDifferenz: -1.689      -------------------
Bias: -3.422                    Node 1: -2.776
                                Node 2: -0.016
Node 3 (Sigmoid)                Node 3: 0.669
----------------                Node 4: -0.251
MarktwertDifferenz: 4.268       Threshold: -1.143
GegentoreDifferenz: -0.755
Bias: 1.143                     Class '2' (Sigmoid)
                                -------------------
Node 4 (Sigmoid)                Node 1: 1.439
----------------                Node 2: -2.941
MarktwertDifferenz: 3.263       Node 3: -2.271
GegentoreDifferenz: -1.932      Node 4: -2.657
Bias: -4.085                    Threshold: 0.369
```

Abbildung 16: Daten des Ergebnismodells des künstlichen neuronalen Netzes

A.3 Ergebnisse der logistischen Regression

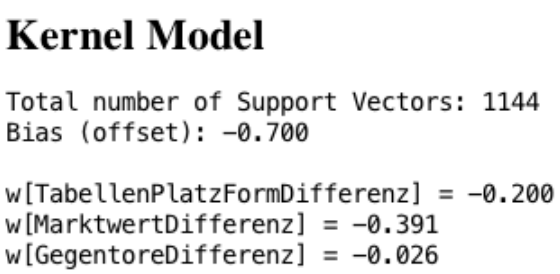


Abbildung 17: Ergebnismodell der logistischen Regression - ’1’ vs. ’X’

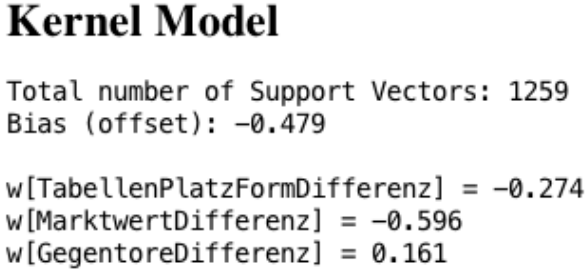


Abbildung 18: Ergebnismodell der logistischen Regression - ’1’ vs. ’2’

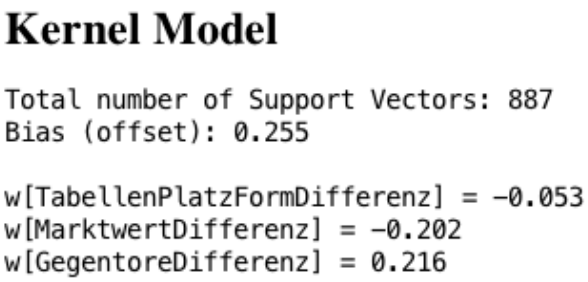


Abbildung 19: Ergebnismodell der logistischen Regression - ’X’ vs. ’2’

Literaturverzeichnis

[Aggarwal 2015] Aggarwal, C. C.: *Data Mining: The Textbook*. New York : Springer Verlag, 2015

[Alpar und Niedereichholz 2000] Alpar, P. ; Niedereichholz, J: *Data Mining im praktischen Einsatz: Verfahren und Anwendungsfälle für Marketing, Vertrieb, Controlling und Kundenunterstützung*. Wiesbaden : Vieweg+Teubner Verlag, 2000

[Anderson und Sally 2013] Anderson, C. ; Sally, D.: *The Numbers Game: Why Everything You Know About Football is Wrong*. London : Penguin Books Limited, 2013

[Anonym 2005] Anonym: *Basiswissen und Wettstrategien*. 2005. – URL `http://wettstrategie.npage.de/basiswissen.html`. – Zugriffsdatum: 31.07.2016

[Bauer und Günzel 2009] Bauer, A. ; Günzel, H.: *Data-Warehouse-Systeme - Architektur, Entwicklung, Anwendung*. 3. überarbeitete und aktualisierte Auflage. Köln : Dpunkt-Verlag, 2009

[Bauer 1998] Bauer, G.: Spiele richtig analysieren – Siege erfolgreich vorbereiten. In: *Fussballtraining* 5 (1998), S. 12–17

[Biermann 2011] Biermann, C.: *Die Fußball-Matrix - Auf der Suche nach dem perfekten Spiel*. Köln : Kiepenheuer & Witsch, 2011

[Biermann 2015] Biermann, C.: *Midtjyllands Revolution - Moneyball im Niemandsland*. 2015. – URL `http://www.11freunde.de/artikel/midtjyllands-revolution`. – Zugriffsdatum: 14.05.2016

[Bramer 2007] Bramer, M.: *Principles of Data Mining*. London : Springer Science & Business Media, 2007

[Buursma 2015] Buursma, D.: Predicting sports events from past results. In: *International Journal of Computer Applications* 132 (2015), S. 8 – 11

[Capello 2010a] Capello, F.: *Capello Index scores – A couple of notes about the scoring system*. 2010. – URL `http://web.archive.org/web/20100917153507/http://www.capelloindex.com/en/news-detail.aspx?id=72ef9154-6e58-4576-8bed-ded78158d84e`. – Zugriffsdatum: 24.05.2016

[Capello 2010b] Capello, F.: *Rankings*. 2010. – URL `http://web.archive.org/web/20100831115032/http://www.capelloindex.com/en/ranking-detail.aspx?chmp_id=313`. – Zugriffsdatum: 24.05.2016

[Carpita u. a. 2015] Carpita, M. ; Sandri, M. ; Simonetto, A. ; Zuccolotto, P.: Discovering the Drivers of Football Match Outcomes with Data Mining. In: *Quality Technology and Quantitative Management* 12 (2015), S. 561 – 577

[Carron u. a. 2005] Carron, A.V. ; Loughhead, T.M. ; Bray, S.R.: The home advantage in sport competitions: Courneya and Carron's (1992) conceptual framework a decade later. In: *Journal of Sports Sciences* 23 (2005), S. 395 — 407

[Chamoni 1999] Chamoni, P.: Ausgewählte Verfahren des Data Mining. In: Chamoni, P., Gluchowski, P. (Hrsg.): *Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining.* Berlin : Springer, 1999, S. 355–373

[Chapman u. a. 1999] Chapman, P. ; Clinton, J. ; Kerber, R. ; Khabaza, T ; Reinartz, T. ; Shearer, C. ; Wirth, R.: *CRISP-DM 1.0 - Step-by-step Data Mining Guide.* Kopenhagen : SPSS, 1999

[Cleve und Lämmel 2014] Cleve, J. ; Lämmel, U.: *Data Mining.* Oldenbourg : De Gruyter, 2014

[Czwalina 1988] Czwalina, C.: *Systematische Spielerbeobachtung in den Sportspielen.* Hamburg : Verlag Hofmann, 1988

[Deshpande 2012] Deshpande, B.: *2 ways of using Naive Bayes classification for numeric attributes.* 2012. – URL `http://www.simafore.com/blog/bid/107702/2-ways-of-using-Naive-Bayes-classification-for-numeric-attributes`. – Zugriffsdatum: 19.09.2016

[Deutsche Fußball Liga GmbH 2012] Deutsche Fussball Liga GmbH: *Pressemitteilung: Ligavorstand beschließt Verteiler-Schlüssel.* 2012. – URL `http://www.fernsehgelder.de/pdf/20121114-DFL.pdf`. – Zugriffsdatum: 26.03.2016

[Deutsche Fußball Liga GmbH 2015] Deutsche Fussball Liga GmbH: *Tabelle - 34. Spieltag.* 2015. – URL `http://www.bundesliga.de/de/liga/tabelle/`. – Zugriffsdatum: 25.06.2016

[Deutsche Fußball Liga GmbH 2016] Deutsche Fussball Liga GmbH: *Bundesliga Report 2016.* Frankfurt : ohne Verlag, 2016

[Deutscher Fußball-Bund 2015] Deutscher Fussball-Bund: *Fußball-Regeln 2015/2016.* Frankfurt : ohne Verlag, 2015

[Di Salvo u. a. 2006] Di Salvo, V. ; Collins, A. ; McNeill, B. ; Cardinale, M.: Validation of Prozone: A new video-based performance analysis system. In: *International Journal of Performance Analysis in Sport* (2006), S. 108–119

[Dorschel 2015] Dorschel, J.: *Praxishandbuch Big Data - Wirtschaft – Recht – Technik.* Berlin : Springer Verlag, 2015

[Fayyad u. a. 1996] Fayyad, U. M. ; Piatetsky-Shapiro, G. ; Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17 (1996), S. 37–54

[FC Bayern München AG 2015] FC Bayern München AG: *Die neue SAP-Station „Spielanalyse“*. 2015. – URL https://fcbayern.com/erlebniswelt/de/news/2015/04/sap. – Zugriffsdatum: 14.05.2016

[Flinders 2002] Flinders, K.: *Football injuries are rocket science*. 2002. – URL http://web.archive.org/web/20080930184834/http://www.vnunet.com/vnunet/news/2120386/football-injuries-rocket-science. – Zugriffsdatum: 23.05.2016

[Fraunhofer Institut für Integrierte Schaltungen 2016] Fraunhofer Institut für Integrierte Schaltungen: *RedFIR*. 2016. – URL http://www.iis.fraunhofer.de/de/ff/lok/proj/redfir.html. – Zugriffsdatum: 03.07.2016

[Gomes u. a. 2015] Gomes, J. ; Portela, F. ; Santos, M.F.: Decision Support System for predicting Football Game result. In: *19th International Conference on Circuits, Systems, Communication and Computers* 32 (2015), S. 348 – 353

[Grabiner 1994] Grabiner, D.: *The Sabermetric Manifesto*. 1994. – URL http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/. – Zugriffsdatum: 03.05.2016

[Grund 2013] Grund, R.: Wertvolles Wissen entdecken und Risiken vermeiden - Data Mining in der Praxis. In: Deggendorfer Forum zur digitalen Datenanalyse e.V. (Hrsg.): *Big Data - Systeme und Prüfung*. Berlin : Erich Schmidt Verlag, 2013, S. 29–44

[Grüling 2015] Grüling, B.: *Spielanalysen - Wie Big Data den Profi-Fußball verändert*. 2015. – URL http://www.welt.de/wissenschaft/article143136567/Wie-Big-Data-den-Profi-Fussball-veraendert.html. – Zugriffsdatum: 25.06.2016

[Hamilton 2010] Hamilton, H.: *What's so wrong with the Capello Index?* 2010. – URL http://www.soccermetrics.net/player-performance/capello-index-discussion. – Zugriffsdatum: 24.05.2016

[Hedley 2016] Hedley, J.: *jsoup: Java HTML Parser*. 2016. – URL https://www.jsoup.org. – Zugriffsdatum: 02.08.2016

[Heuer und Rubner 2012] Heuer, A. ; Rubner, O.: Towards the perfect prediction of soccer matches. (2012)

[Heuer und Rubner 2013] Heuer, A. ; Rubner, O.: Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer. (2013)

[Heuer 2013] Heuer, Andreas: *Der perfekte Tipp - Statistik des Fußballspiels*. New York : John Wiley & Sons, 2013

[Hippner und Wilde 2009] Hippner, H ; Wilde, K: Data Mining im CRM. In: Helmke, S. ; Uebel, M. ; Dangelmaier, W. (Hrsg.): *Effektives Customer Relationship Management - Instrumente - Einführungskonzepte - Organisation.* 4. Auflage. Berlin : Springer Verlag, 2009, S. 205–225

[Inmon 2005] Inmon, W.H.: *Building the Data Warehouse.* New York : John Wiley & Sons, Inc., 2005

[Kemper u. a. 2010] Kemper, H.-G. ; Baars, H. ; Mehanna, W.: *Business Intelligence - Grundlagen und praktische Anwendungen - Eine Einführung in die IT-basierte Managementunterstützung.* 3. Auflage. Wiesbaden : Vieweg+Teubner Verlag, 2010

[Kjäll 2015] Kjäll, A.: *Midtjylland: Meet the men behind Moneyball FC.* 2015. – URL `http://www.fourfourtwo.com/features/midtjylland-meet-men-behind-moneyball-fc`. – Zugriffsdatum: 25.05.2016

[Knobloch und Weidner 2000] Knobloch, B. ; Weidner, J.: Eine kritische Betrachtung von Data Mining-Prozessen – Ablauf, Effizienz und Unterstützungspotenziale. In: Jung, R., Winter, R. (Hrsg.): *Data Warehousing 2000: Methoden, Anwendungen, Strategien.* Heidelberg : Physica-Verlag HD, 2000, S. 345–365

[Kroos 2016] Kroos, T.: *Toni Kroos: Statistik.* 2016. – URL `http://www.toni-kroos.com/de/stats`. – Zugriffsdatum: 22.05.2016

[Kuper 2011] Kuper, S.: *The Football Men - Up Close with the Giants of the Modern Game.* New York : Simon and Schuster, 2011

[Kuper und Szymanski 2012] Kuper, Simon ; Szymanski, Stefan: *Soccernomics.* 2. Auflage. London : HarperCollins UK, 2012

[Köppen u. a. 2012] Köppen, V. ; Saake, G. ; Sattler, K.-U.: *Data Warehouse Technologien.* Heidelberg : MITP, 2012

[Küppers 1998] Küppers, B.: *Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld.* Frankfurt am Main : Lang, 1998

[Lago u. a. 2011] Lago, C. ; Lago, J. ; Rey, E.: Differences in performance indicators between winning and losing teams in the UEFA Champions League. In: *Journal of Human Kinetics* Bd. 27. 2011, S. 135–146

[Lames 1994] Lames, M.: *Systematische Spielbeobachtung.* Greven : Philippka, 1994

[Lames 1999] Lames, M.: Fußball - Ein Chaosspiel? In: Janssen, J.-P.; Wilhelm, A.; Wegner, M. (Hrsg.): *Empirische Forschung im Sportspiel.* Kiel : Institut für Sportwissenschaft Kiel, 1999, S. 141–156

[Lange 1997] Lange, P.: Tore sind kein Zufallsprodukt! Taktische Lösungsmöglichkeiten für das Angriffsspiel: Mit Spielzügen Torchancen vorbereiten! In: *Fussballtraining* Bd. 15. 1997, S. 35–41

[Laub und Merx 2013] Laub, M. ; Merx, S.: *Exklusivstudie: Was der Bundesligaabstieg wirtschaftlich bedeutet.* 2013. – URL `http://www.jp4sport.biz/archive/5716/exklusivstudie-was-der-bundesligaabstieg-wirtschaftlich-bedeutet/`. – Zugriffsdatum: 26.03.2016

[Leser 2007] Leser, R.: *Computerunterstützte Sportspielanalyse im Fußball - Methoden für den praxisgerechten Einsatz.* Saarbrücken : VDM Verlag Dr. Müller, 2007

[Lewis 2004] Lewis, M.: *Moneyball: The Art of Winning an Unfair Game.* New York : W. W. Norton, 2004

[Loy 2008] Loy, R.: *Das Lexikon der Fußballirrtümer.* München : C. Bertelsmann Verlag, 2008

[Mangat 2015] Mangat, R.: *Is Brentford's Analytics Revolution the Future of Football in England?* 2015. – URL `http://thesefootballtimes.co/2015/08/31/is-brentfords-analytics-revolution-the-future-of-football-in-england/`. – Zugriffsdatum: 14.05.2016

[Matthäus und Schulze 2015] Matthäus, W.G. ; Schulze, J.: *Statistik mit Excel: Beschreibende Statistik für jedermann.* Wiesbaden : Vieweg+Teubner Verlag, 2015

[Medeiros 2014] Medeiros, J.: *The winning formula: data analytics has become the latest tool keeping football teams one step ahead.* 2014. – URL `http://www.wired.co.uk/article/the-winning-formula`. – Zugriffsdatum: 26.05.2016

[Nopp 2012] Nopp, S.: *Direkt- versus Ballbesitzspiel: erfolgreiche mannschaftstaktische Angriffsdeterminanten auf nationalem und internationalem Niveau im Sportspiel Fußball*, Deutsche Sporthochschule Köln, Dissertation, 2012

[Oakland Athletics 2014] Oakland Athletics: *Postseason Results.* 2014. – URL `http://oakland.athletics.mlb.com/oak/history/postseason_results.jsp`. – Zugriffsdatum: 05.05.2016

[Oberstone 2011] Oberstone, J.: Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008-2009 Season. In: *Journal of Quantitative Analysis in Sports* Bd. 7. 2011, S. 1–18

[Ofoghi u. a. 2013] Ofoghi, B. ; Zeleznikow, J. ; MacMahon, C. ; Raab, M.: Data Mining in Elite Sports: A Review and a Framework. In: Wood, T.M. (Hrsg.): *Measurement in Physical Education and Exercise Science* Bd. 17. London : Taylor & Francis, 2013, S. 171–186

[Opta Sports 2013a] Opta Sports: *Bundesliga Saison 2013/14*. 2013. – URL http://www.optasports.de/showcase-pages/bundesliga-20132014.aspx. – Zugriffsdatum: 03.07.2016

[Opta Sports 2013b] Opta Sports: *Opta's Event Definitions*. 2013. – URL http://optasports.com/news-area/blog-optas-event-definitions.aspx. – Zugriffsdatum: 25.06.2016

[Opta Sports 2016a] Opta Sports: *Die Datenerfassung*. 2016. – URL http://www.optasports.de/de/über-uns/so-arbeiten-wir/the-data-collection-process.aspx. – Zugriffsdatum: 03.07.2016

[Opta Sports 2016b] Opta Sports: *Kunden & Partner*. 2016. – URL http://www.optasports.de/über-uns/kunden-partner.aspx. – Zugriffsdatum: 03.07.2016

[Opta Sports 2016c] Opta Sports: *Opta data feeds overview*. 2016. – URL http://www.optasports.com/services/media/data-feeds/opta-data-feeds-overview.aspx. – Zugriffsdatum: 03.07.2016

[Owramipur u. a. 2013] Owramipur, F. ; Eskandarian, P. ; Mozneb, F.S.: Football Result Prediction with Bayesian Network in Spanish League – Barcelona Team. In: *International Journal of Computer Theory and Engineering* 5 (2013), S. 812 – 815

[Quitzau und Völpel 2009] Quitzau, J. ; Völpel, H.: *Der Faktor Zufall im Fußball: Eine empirische Untersuchung für die Saison 2007/08*. Hamburg : Hamburgisches WeltWirtschaftsInstitut (HWWI), 2009

[Rapid-I 2010] Rapid-I: *Rapidminer - Benutzerhandbuch*. o.O. : o.V., 2010

[RapidMiner 2014] RapidMiner: *Rapidminer Studio - Manual*. o.O. : o.V., 2014

[RapidMiner 2016a] RapidMiner: *Neural Net – RapidMiner Studio Core*. 2016. – URL http://docs.rapidminer.com/studio/operators/modeling/predictive/neural_nets/neural_net.html. – Zugriffsdatum: 20.09.2016

[RapidMiner 2016b] RapidMiner: *Optimize Selection – RapidMiner Studio Core*. 2016. – URL http://docs.rapidminer.com/studio/operators/modeling/optimization/feature_selection/optimize_selection.html. – Zugriffsdatum: 12.09.2016

[RapidMiner 2016c] RapidMiner: *X-Validation – RapidMiner Studio Core*. 2016. – URL http://docs.rapidminer.com/studio/operators/validation/x_validation.html. – Zugriffsdatum: 11.09.2016

[Reisenweber 2016] Reisenweber, B.: *What is the pricing model and costs of OPTA Services, sports data company?* 2016. – URL https://www.quora.com/

What-is-the-pricing-model-and-costs-of-OPTA-Services-sports-data-company/answer/Bob-Reisenweber. – Zugriffsdatum: 03.07.2016

[Rotshtein u. a. 2005] Rotshtein, A. P. ; Posner, M. ; Rakityanskaya, A. B.: Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning. In: *Cybernetics and System Analysis* 41 (2005), S. 619–630

[Runkler 2015] Runkler, T.: *Data Mining - Modelle und Algorithmen intelligenter Datenanalyse*. 2. Auflage. Berlin : Springer Verlag, 2015

[Russell und Norvig 2004] Russell, S. ; Norvig, P: *Künstliche Intelligenz. Ein moderner Ansatz*. 4. Auflage. München : Pearson Studium, 2004

[SAP SE 2014] SAP SE: *WM in Brasilien: Spickzettel war gestern*. 2014. – URL http://news.sap.com/germany/2014/06/17/wm-brasilien-spickzettel-war-gestern-jetzt-gibt-es-apps-zur-datenauswertung/. – Zugriffsdatum: 14.05.2016

[Schumaker u. a. 2010] Schumaker, R. ; Solieman, O. ; Chen, H.: *Sports Data Mining*. Berlin : Springer Verlag, 2010

[Sharafi 2013] Sharafi, A.: *Knowledge Discovery in Databases - Eine Analyse des Änderungsmanagements in der Produktentwicklung*. Berlin : Springer Verlag, 2013

[Sharma und Osei-Bryson 2009] Sharma, S. ; Osei-Bryson, K.: Framework for formal implementation of the business understanding phase of data mining projects. In: *Expert Systems with Applications* 36 (2009), S. 4114–4124

[Skinner und Freeman 2009] Skinner, G.K. ; Freeman, G.H.: Are soccer matches badly designed experiments? In: *Journal of Applied Statistics* 36 (2009), S. 1087–1095

[Steinbrecher und Schumann 2015] Steinbrecher, M. ; Schumann, R.: *Update - Warum die Datenrevolution uns alle betrifft*. Frankfurt am Main : Campus Verlag, 2015

[Sumpter 2016] Sumpter, David: *Soccermatics - Mathematical Adventures in the Beautiful Game*. New York : Bloomsbury Publishing, 2016

[Tan u. a. 2006] Tan, P.N. ; Steinbach, M. ; Kumar, V.: *Introduction to Data Mining*. Boston : Pearson Addison Wesley, 2006

[Transfermarkt.de 2016a] Transfermarkt.de: *FC Barcelona – Historische Platzierungen*. 2016. – URL http://www.transfermarkt.de/fc-barcelona/platzierungen/verein/131. – Zugriffsdatum: 31.07.2016

[Transfermarkt.de 2016b] Transfermarkt.de: *Transfermarkt Community - Marktwertanalyse*. 2016. – URL http://www.transfermarkt.de/intern/community. – Zugriffsdatum: 02.08.2016

[TSG 1899 Hoffenheim 2012] TSG 1899 Hoffenheim: *1899 profitiert von neuer SAP-Technologie*. 2012. – URL http://www.achtzehn99.de/newsarchiv-2/newsarchiv-2013/november-2013/1899-profitiert-von-neuer-sap-technologie/. – Zugriffsdatum: 25.05.2016

[UEFA 2014] UEFA: *Frühwarnsystem gegen Wettbetrug*. 2014. – URL http://de.uefa.org/protecting-the-game/integrity/betting-fraud-detection-system/index.html. – Zugriffsdatum: 22.05.2016

[Ulmer und Fernandez 2014] Ulmer, B. ; Fernandez, M.: Predicting Soccer Match Results in the English Premier League. (2014)

[Universität Münster 2016] Universität Münster: *Sportstatistik*. 2016. – URL https://www.uni-muenster.de/Chemie.pc/heuer/sport_statistics.html. – Zugriffsdatum: 31.07.2016

[Vossen 2008] Vossen, G.: *Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme*. 5. Auflage. Deutschland : Oldenbourg, 2008

[Weiss und Indurkhya 1998] Weiss, S.M. ; Indurkhya, N.: *Predictive Data Mining - A Practical Guide*. San Francisco : Morgan Kaufmann, 1998

[Winkler 2000] Winkler, W: Analyse von Fußballspielen mit Video- und Computerhilfe. In: *Computer- und Medieneinsatz im Fußball - 13. Jahrestagung der Dvs-Kommission Fußball*. 2000, S. 63–76

[Witten und Frank 2001] Witten, I.H. ; Frank, E.: *Data mining - praktische Werkzeuge und Techniken für das maschinelle Lernen*. München : Hanser, 2001

[Wu u. a. 2014] Wu, X. ; Zhu, X. ; Wu, G. ; Ding, W.: Data mining with big data. In: *IEEE Transactions on Knowledge and Data Engineering* 26 (2014), Nr. 1, S. 97–107

[ZDF Sport 2015] ZDF Sport: *Bayern München - Hamburger SV: Spieldaten*. 2015. – URL http://www.zdfsport.de/ZDF/zdfportal/cacheable/6042950/2/4674/796f99?tabellengruppenId=34230050&userFilterNames=f1&userFilterValues=489&userFilterNames=f2&userFilterValues=1985665&action=setActionData. – Zugriffsdatum: 25.06.2016

Erklärung gemäß § 21 Abs 9 und § 23 Abs. 2 der Prüfungsordnung der Universitäten Hohenheim und Stuttgart für den Masterstudiengang Wirtschaftsinformatik

Hiermit erkläre ich, dass ich die Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderen fremden Äußerungen entnommen wurden, sind als solche einzeln kenntlich gemacht.

Die Masterarbeit habe ich noch nicht in einem anderen Studiengang als Prüfungsleistung verwendet.

Des Weiteren erkläre ich, dass mir weder an den Universitäten Hohenheim und Stuttgart noch an einer anderen wissenschaftlichen Hochschule bereits ein Thema zur Bearbeitung als Masterarbeit oder als vergleichbare Arbeit in einem gleichwertigen Studiengang vergeben worden ist.

Stuttgart-Hohenheim, den

Unterschrift