Lehrstuhl Angewandte Informatik IV
Datenbanken und Informationssysteme
Prof. Dr.-Ing. Stefan Jablonski

Institut für Angewandte Informatik
Fakultät für Mathematik, Physik und Informatik
Universität Bayreuth

Bachelor Seminar

Marcel Fraas

4. Februar, 2017
Version: Draft

Universität Bayreuth

Fakultät Mathematik, Physik, Informatik
Institut für Informatik
Lehrstuhl für Angewandte Informatik IV

Data Mining Frameworks

Bachelor Seminar

Marcel Fraas

1. Reviewer Prof. Dr.-Ing. Stefan Jablonski
Fakultät Mathematik, Physik, Informatik
Universität Bayreuth

2. Reviewer Dr. Stefan Schönig
Fakultät Mathematik, Physik, Informatik
Universität Bayreuth

Supervisors Stefan Schönig and Lars Ackermann

4. Februar, 2017

Marcel Fraas
Bachelor Seminar
Data Mining Frameworks, 4. Februar, 2017
Reviewers: Prof. Dr.-Ing. Stefan Jablonski and Dr. Stefan Schönig
Supervisors: Stefan Schönig and Lars Ackermann

Universität Bayreuth
Lehrstuhl für Angewandte Informatik IV
Institut für Informatik
Fakultät Mathematik, Physik, Informatik
Universitätsstrasse 30
95447 Bayreuth
Germany

Inhaltsverzeichnis

1 Grundlagen zu Data Mining

> *Information is not knowledge.*
>
> **— Albert Einstein**
> (Theoretischer Physiker)

Der Begriff Data Mining bezeichnet zunächst einmal das Sammeln, Verarbeiten und Analysieren von Daten und den damit verbundenen Informationsgewinn. Da allerdings in der echten Welt eine große Bandbreite an Anwendungen und Problemfeldern existiert, versteht man unter dem „minen von Daten“ ein sehr weit gefächertes Feld an Methoden zur Datenverarbeitung.

Data Mining hat in unserem Alltag längst Einzug gefunden, meistens bemerken wir dies jedoch gar nicht. Nutzen wir beim Einkaufen beispielsweise eine Bonuspunkte-Karte, sind in sozialen Medien aktiv oder stehen auf dem Weg zur Arbeit im Stau generieren wir eine Unmenge an Daten. Diese werden von Unternehmen gesammelt und anschließend ausgewertet. Innerhalb dieser Ansammlung an Datensätzen finden sich Informationen über unsere Gewohnheiten, unsere Interessen und über unser Verhalten.
Data Mining hilft uns eben jene Informationen interpretierbar zu machen und somit besser zu verstehen wie Menschen mit ihrer Umwelt interagieren.

Gleichzeitig muss man allerdings auch den Aspekt des Datenschutzes beachten. Reicht das sammeln von Daten zu weit in die Privatsphäre eines einzelnen, kann dies schnell zum Missbrauch dieser Informationen führen.

Dass das Thema Data Mining kontrovers ist zeigt auch der Artikel „How Companies Learn Your Secrets“ aus dem New York Times Magazine. [Duh12] Hier wollte eine amerikanische Supermarktkette das Kaufverhalten ihrer Kunden untersuchen. Um dies zu bewerkstelligen wurde den Kunden zunächst eine Identifikationsnummer zugewiesen, sowie Namen, Kreditkarteninformationen und Email-Adresse gespeichert. Unter Einbezug weiterer externer Datenquellen zur Demografie konnten einige interessante Beobachtungen gemacht werden. So konnte die Supermarktkette beispielsweise die Schwangerschaft von Frauen anhand der Einkäufe erkennen und hat im Zuge dessen festgestellt, dass schwangere Frauen im zweiten Trimester ihrer Schwangerschaft vermehrt geruchlose Lotionen kaufen. Außerdem werden

innerhalb der ersten 20 Wochen der Schwangerschaft häufiger Zusatzstoffe wie Kalzium, Magnesium und Zink erworben. Nähert sich der Tag der Entbindung werden zunehmend geruchlose Seife und extra große Wattepads in Verbindung mit Desinfektionsmittel und Waschlappen gekauft.
Mit diesen Informationen war es möglich einen sog. „pregnancy prediction score“ zu errechnen welcher dazu genutzt wurde gezielt Werbung in Form von Gutscheinen zu bestimmten Zeiten der Schwangerschaft zu verschicken.
Der Artikel beschreibt einen Vorfall bei dem ein wütender Mann in eine der Filialen der Supermarktkette kam und sich beschwerte, dass seine Tochter Gutscheine für Babykleidung und Krippen bekam. Der Mann wolle nicht, dass seine Tochter von der Supermarktkette dazu ermutigt werde schwanger zu werden. Als der Manager der Filiale später bei dem Vater anrief um sich zu entschuldigen wurde ihm von diesem mitgeteilt, dass die Tochter tatsächlich schwanger sei, was der Vater jedoch zum Zeitpunkt seiner Beschwerde nicht wusste.

Dieses Beispiel zeigt, dass es wichtig ist präzise abzuwägen wie genau eine Analyse der Daten sein sollte.

2 Der Data Mining Prozess

Der Prozess des Data Minings lässt sich grundsätzlich durch die folgenden Phasen beschreiben:

1. Sammeln von Daten:
   Für das Sammeln von Daten ist unter Umständen spezielle Hardware, bspw. Sensoren, händische Arbeit wie das Sammeln von Umfragebögen oder Software in Form einer Webanwendung mit auszufüllenden Formularfeldern, notwendig. Auch wenn diese Phase sehr Anwendungsspezifisch ist und oft vom Daten-Analysten nicht beeinflusst werden kann, ist sie ausschlaggebend für das Ergebnis des Data Mining Prozesses.

2. Bereinigen der Daten:
   Oftmals sind die gesammelten Rohdaten aufgrund des Dateiformats oder einer fehlenden Struktur nicht direkt verarbeitbar. Deshalb ist es wichtig, die Daten in ein Format zu bringen, welches von Data Mining Algorithmen gelesen werden kann.

3. Analyse der Daten:
   Der letzte Schritt ist die Daten analytisch zu verarbeiten und Methoden zu entwickeln, um diese nutzbar zu machen. Dies geschieht durch das Anwenden von bestimmten Data Mining Algorithmen, welche für die jeweilige Aufgabenstellung angepasst werden müssen.

Um einen Standard für einen solchen Data Mining Vorgang zu etablieren, haben einige große Unternehmen wie Automobilhersteller Daimler-Benz, Versicherungs Provider OHRA, Hard- und Software Hersteller NCR Corp. und Statistik-Software Hersteller SPSS Inc. den „CRoss-Industry Standard Process for Data Mining“ (kurz: CRISP-DM) definiert.

2.1 CRISP-DM

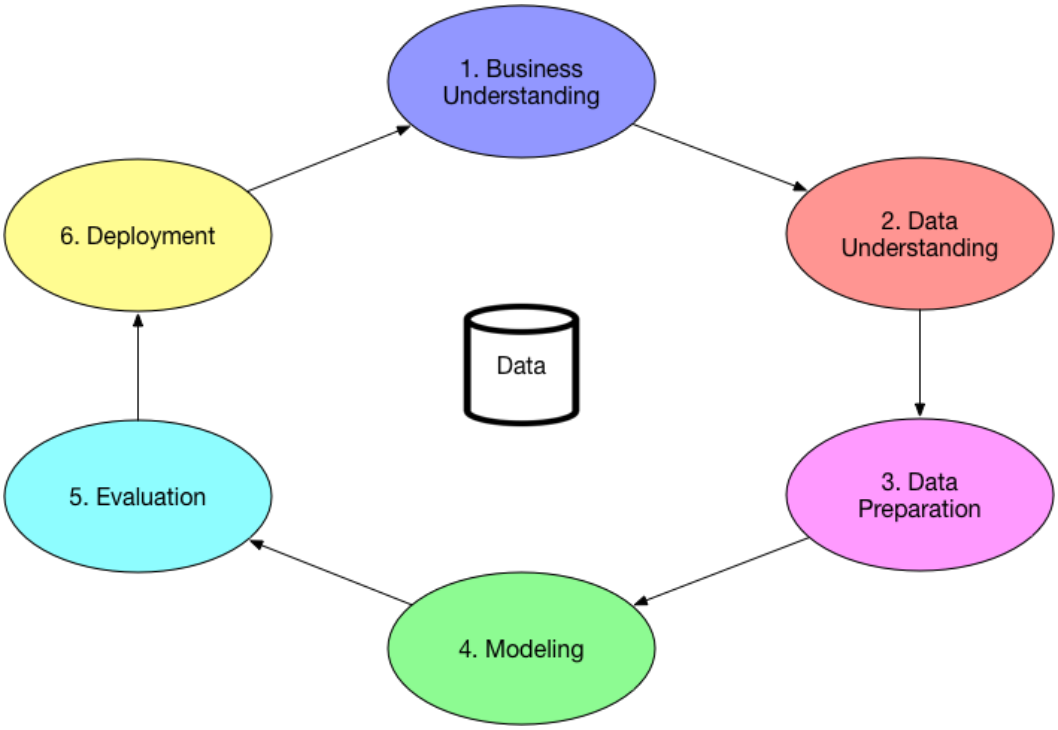

Abbildung 2.1: Konzeptionelles CRISP-DM Modell

2.1.1 Business Understanding

Im ersten Schritt des CRISP-DM Prozesses, dem sog. Business- (oder auch Organizational) Understanding, ist es zunächst wichtig festzulegen was man mit dem Minen von Daten erreichen möchte bzw. welche Informationen von Interesse sind. Hier findet auch eine oft zitierte Textstelle aus Alice im Wunderland Anwendung:

> »Willst du mir wohl sagen, wenn ich bitten darf, welchen Weg ich hier nehmen muß?«
> »Das hängt zum guten Teil davon ab, wohin du gehen willst,« sagte die Katze.
> »Es kommt mir nicht darauf an, wohin –« sagte Alice.
> »Dann kommt es auch nicht darauf an, welchen Weg du nimmst,« sagte die Katze.
> »– wenn ich nur irgendwo hinkomme,« fügte Alice als Erklärung hinzu.
> »O, das wirst du ganz gewiß,« sagte die Katze, »wenn du nur lange genug gehest.«

In Bezug auf das Thema Data Mining bedeutet das, dass es keine Rolle spielt wie lange man Daten Mined, wenn man nicht definiert hat welche Informationen man

gerne hätte. Es müssen also zunächst Fragen festgelegt werden, welche durch das Data Mining beantwortet werden sollen. Beispielsweise möchte man gerne wissen warum sich Kunden so sehr beschweren, wie man die Profit-Spanne seiner Produkte vergrößert oder wie man Fehler bei der Herstellung antizipieren kann.

2.1.2 Data Understanding

Man muss zunächst zwischen zwei Systemen unterscheiden.

- OLTP - Online Transaction Processing System:
  Als OLTP werden die meisten relationalen Datenbank Systeme bezeichnet. Sie sind ausgelegt für eine große Anzahl an „Reads“ und „Writes“. Ein Beispiel hierfür ist der Kassiervorgang im Supermarkt, bei dem in einer kurzen Zeit viele Gegenstände per Barcode registriert werden müssen. Diese Systeme sind durch die Normalisierung nicht besonders für die Analyse geeignet, da mitunter sehr viele Joins ausgeführt werden müssen.

- OLAP - Online Analytical Processing System:
  Hat man die Daten in Form eines Data Warehouses in denormalisierter Form vorliegen, spricht man von einem OLAP. Wie der Name schon verrät eignen sich diese Datenbank Systeme zur Analyse der Daten, da hier die Datensätze zu einer geringen Zahl an Tabellen zusammengefasst wurden. Zu Beachten ist dabei allerdings, dass durch den Vorgang der Denormalisierung auch Redundanz auftritt und daher mehr Speicherplatz benötigt wird.

Schließlich muss man auch die Daten selbst unterscheiden. Hier existieren zwei Typen mit welchen man Data Mining betreiben kann.

- Operational Data:
  Diese Daten stammen aus Systemen, welche auf Transaktionen basieren. Dies können beispielsweise Daten aus Online-Bestellungen, Check-In Informationen am Flughafen oder anderen alltäglichen Aktivitäten sein. Allerdings sind diese Daten auch sehr detailliert und könnten daher unter Umständen die Privatsphäre verletzen.

- Organizational Data:
  Hierbei handelt es sich um Daten welche anonymisiert und zusammengefasst wurden. Dadurch wird sowohl die Privatsphäre des Einzelnen geschützt als auch die Möglichkeit geboten effizient Informationen wie bspw. Trends zu erkennen.

Wichtig ist hier die Zuverlässigkeit und Genauigkeit der Daten zu überprüfen, denn Entscheidungen basierend auf ungenauen Daten sind auch entsprechend ungenau.

2.1.3 Data Preparation

In diesem Schritt müssen die gesammelten Rohdaten verarbeitbar gemacht werden. Beispielsweise müssen verfälschte bzw. für die Analyse unwichtige Daten herausgefiltert, Attribute und deren Typen transformiert oder generell Datensätze bereinigt werden. Letztendlich müssen die Daten so aufbereitet werden, dass die Data Mining Algorithmen im nächsten Schritt damit arbeiten können. Dies ist daher auch der aufwendigste Schritt im kompletten CRISP-DM Prozess.

2.1.4 Modeling

Die Data Mining Modelle kann man wiederum in 2 Arten unterteilen.

- Deskriptive Modelle (engl.: „descriptive Model“):
  Diese Art von Modell trifft zwar keine Vorhersage über zukünftige Werte, kann aber Informationen über die immanente Struktur der Daten und Relationen liefern. Ein Beispiel hierfür ist die sog. Korrelationsmatrix

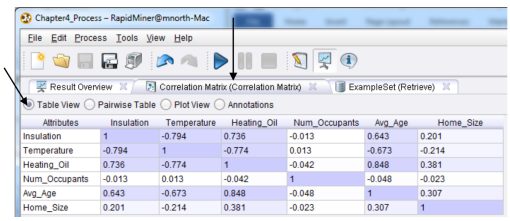

Abbildung 2.2: Ergebnis einer Korrelationsmatrix [Nor12]

  Korrelation bezeichnet eine statistische Aussage darüber wie stark Beziehungen zwischen Attributen in einem Datensatz sind.

- Vorhersagende Modelle (engl.: „predictive Model“):
  Wie sich anhand des Namens vermuten lässt, sagt ein vorhersagendes Modell einen bestimmten oder mehrere Werte voraus. Entscheidungsbäume (engl.: Decision Trees) sind ein Beispiel für ein solches vorhersagendes Modell.

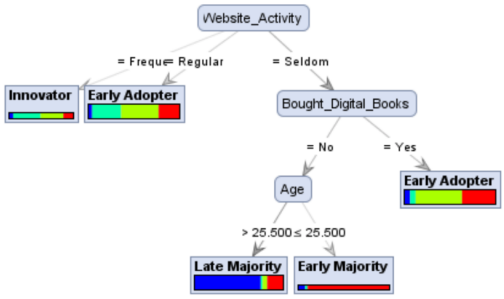

Abbildung 2.3: Ergebnis eines Entscheidungsbaums [Nor12]

> Entscheidungsbäume sind eine grafische Darstellung von hierarchisch aufeinanderfolgenden Entscheidungen.

Prinzipiell versucht man beim Modeling ein Modell zu erschaffen (bzw. zu „trainieren“), welches die echte Weit so gut wie möglich repräsentiert. Man tut dies, um damit entweder fundierte Aussagen über die Zukunft treffen zu können oder bisher unbekannte Informationen über den aktuellen Zustand zu erhalten.

2.1.5 Evaluation

Hat man ein Modell trainiert und ein entsprechendes Ergebnis erhalten muss man im nächsten Schritt prüfen, ob dieses auch sinnvoll ist. Analysen können, bspw. aufgrund von falschen Parametern im Algorithmus oder durch ungenaue Daten, fehlerhaft sein. Ist dies der Fall muss entweder das Modell entsprechend angepasst werden bzw. die Daten im Schritt „Data Preparation“ weiter bearbeitet werden. Außerdem muss das Ergebnis auf Relevanz und das Model somit auf Aussagekraft geprüft werden. Basieren Ergebnisse auf einer geringen Anzahl an Datensätzen, kann dies zu fälschlichen Annahmen bzw. unzutreffenden Informationen führen.

2.1.6 Deployment

Ist man mit seinem Modell und dessen Aussagekraft zufrieden, kann man den Prozess automatisieren. Dies geschieht durch die Implementierung in ein (wahrscheinlich bereits vorhandenes) Informationssystem. Außerdem kann man mit den gewonnenen Informationen nun Entscheidungen treffen, wobei jedoch folgendes beachtet werden muss:
Korrelation bedeutet nicht Kausalität! Nur weil, bspw. durch eine Korrelationsmatrix, zwischen zwei Attributen eine Korrelation festgestellt wurde, heißt das nicht, dass der Wert eines Attributs der Grund für einen Wert des korrelierenden Attributs ist.

2.2 Alternativen

Die folgende Tabelle zeigt Umfrageergebnisse zu der Frage welche Methode für die Analyse bzw. das Data Mining im Unternehmen der Befragten zum Einsatz komme. Betrachtet man die Ergebnisse tauchen neben einigen spezifischen Prozessen auch der sog. SEMMA Prozess und der KDD Prozess auf.

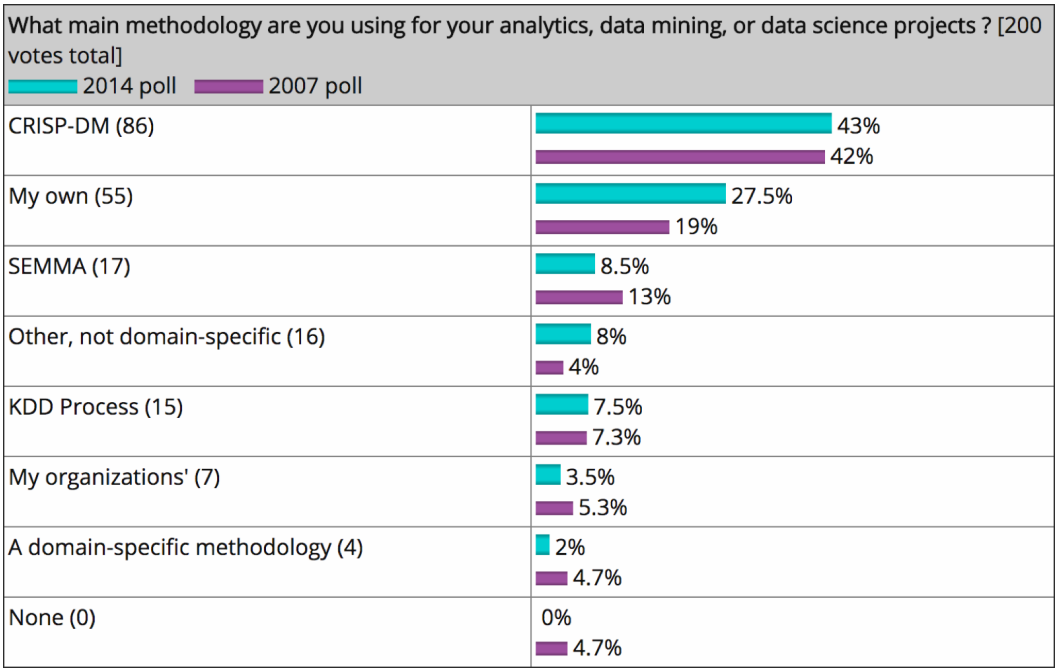

Abbildung 2.4: Umfrage welche DM-Prozesse im Unternehmen zum Einsatz kommen [Kdn]

2.2.1 Der Data Mining Prozess KDD

In Fayyads „Knowledge Discovery in Databases“ [M.96] wird Data Mining als eine der Phasen des Prozesses gesehen, welche zur „Gewinnung von Erkenntnissen“ dienen soll.

Der Prozess umfasst die nachfolgenden Phasen.

1. Selektion (engl.: Selection):
   Mit der Selektion soll ein Ziel-Datensatz definiert werden bzw. eine Menge an Variablen und Beispiel-Datensätzen festgelegt werden welche zur Feststellung neuer Erkenntnisse dienen sollen.

2. Vorverarbeitung (engl.: Pre processing):
   In diesem Schritt sollen die Daten bereinigt werden und dadurch eine gewisse Konsistenz gewährleistet werden.

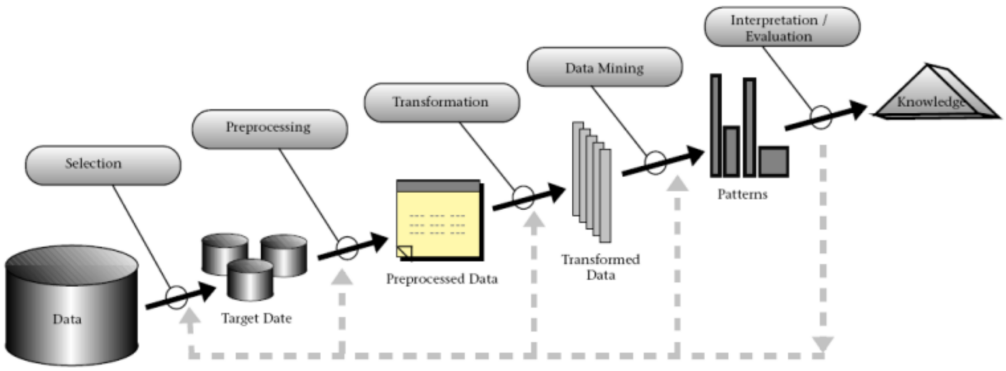

Abbildung 2.5: Schritte die den KDD Prozess zusammensetzen [M.96]

3. Transformation (engl.: Transformation):
   Hier sollen durch anwenden verschiedener Transformationen die Daten auf wesentliche Einträge (vor allem hinsichtlich ihrer Dimension) reduziert werden.

4. Data Mining:
   Mit Data Mining soll nach Mustern innerhalb der Daten gesucht werden und diese entsprechend in eine repräsentative Form gebracht werden.

5. Interpretation / Evaluation:
   Zuletzt sollen die gefundenen Muster hinsichtlich ihrer Aussagekraft evaluiert werden.

In gewisser Hinsicht kann der KDD Prozess mit dem CRISP-DM Prozess verglichen werden:

- Die Phase des „Business Understanding“ (CRISP-DM) korrespondiert mit dem Entwickeln eines Verständnisses des Anwendungsgebiets, dem jeweiligen Vorwissen und dem festlegen eines Ziels für den Endnutzer (KDD)

- Der Schritt „Data Unterstanding“ (CRISP-DM) ist vergleichbar mit einer Kombination aus „Selection“ und „Preprocessing“ (KDD)

- Die „Data Preparation“ (CRISP-DM) kann identifiziert werden mit der „Transformation“ (KDD)

- Die „Modeling“ (CRISP-DM) Phase entspricht in etwa dem „Data Mining“ (KDD)

- „Evaluation“ (CRISP-DM KDD) kann gleichgesetzt werden

- Letztendlich kann „Deployment“ (CRISP-DM) mit der Konsolidierung (KDD) beschrieben werden indem die gewonnenen Erkenntnisse ins System eingebunden werden

2.2.2 Der Data Mining Prozess SEMMA

Das Akronym SEMMA steht für „Sample Explore Modify Model Assess“ und wurde vom SAS Institute entwickelt. Auch wenn der Prozess prinzipiell unabhängig vom gewählten Softwaretool ist, gibt es hier eine Verbindung zur von SAS bereitgestellten Software „SAS Enterprise Miner“.
Die Phasen werden wie folgt beschrieben:

1. Probieren (engl.: Sample):
   In diesem (optionalen) Schritt soll eine Untermenge an Datensätzen ausgewählt werden, welche zwar alle Struktur-relevanten Daten beinhaltet allerdings immer noch klein genug ist um schnell manipuliert werden zu können.

2. Entdecken (engl.: Explore):
   Hier sollen die Datensätze auf unerwartete Trends und Anomalien untersucht werden um ein Verständnis der Daten und der Struktur zu gewinnen.

3. Modifizieren (engl.: Modify):
   Durch Modifikation der Daten, also Selektion und Transformation, sollen diese so kombiniert werden, sodass ein Mining Model ausgewählt bzw. darauf angewendet werden kann.

4. Modellieren (engl.: Model):
   Diese Phase soll dazu dienen, die Daten zu modellieren, d.h. mithilfe einer entsprechenden Software automatisch nach einer Kombination von Datensätzen zu suchen, welche ein gewünschtes Ergebnis vorhersagen.

5. Beurteilen (engl.: Assess):
   Zuletzt soll das Ergebnis beurteilt werden, indem die Genauigkeit und Nützlichkeit der Funde evaluiert werden und abgeschätzt wird wie effizient der Prozess insgesamt ist.

3 Software

4 Datensatz und Beispiel-Modell

4.1 Datensatz

4.1.1 Business Understanding

4.1.2 Data Understanding

4.1.3 Data Preparation

4.1.4 Modeling

4.1.5 Evaluation Deployment

4.2 Umsetzung

4.2.1 Rapidminer Studio

4.2.2 Microsoft Azure Machine Learning Studio

5 Fazit

Literatur

[M.96] Fayyad U. M. *Avdvances in knowledge discovery and data mining*. AAAI Press, 1996 (siehe S. 8, 9).

[Nor12] Dr. Matthew North. *Data Mining for the Masses*. CreateSpace Independent Publishing Platform, 2012 (siehe S. 6, 7).

Websites

[Duh12] Charles Duhigg. *How Companies Learn Your Secrets*. 2012 (siehe S. 1).

[Kdn] *What main methodology are you using for your analytics, data mining, or data science projects ?* 2014 (siehe S. 8).

Abbildungsverzeichnis

Declaration

You can put your declaration here, to declare that you have completed your work solely and only with the help of the references you mentioned.

Bayreuth, 4. Februar, 2017

Marcel Fraas