



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 Data Mining Grundlagen	1
2 Der Data Mining Prozess	3
2.1 CRISP-DM	4
2.1.1 Business Understanding	4
2.1.2 Data Understanding	5
2.1.3 Data Preparation	6
2.1.4 Modeling	6
2.1.5 Evaluation	7
2.1.6 Deployment	7
2.2 Alternativen	8
2.2.1 Der Data Mining Prozess KDD	8
2.2.2 Der Data Mining Prozess SEMMA	10
3 Vorstellung der Software	11
3.1 Rapidminer Studio	11
3.2 Microsoft Azure Machine Learning Studio	13
4 Datensatz und Beispiel-Modell	15
4.1 Datensatz	15
4.1.1 Business Understanding	15
4.1.2 Data Understanding	15
4.1.3 Data Preparation	17
4.1.4 Modeling	17
4.1.5 Evaluation + Deployment	18
4.2 Umsetzung	18
4.2.1 Rapidminer Studio	18
4.2.2 Microsoft Azure Machine Learning Studio	19
5 Fazit	23
Literatur	27

Data Mining Grundlagen

„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.

Der Data Mining Prozess

Der Prozess des Data Minings lässt sich grundsätzlich durch die folgenden Phasen beschreiben [Nor12]:

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 Firmen wie Automobilhersteller Daimler-Benz, Versicherungsunternehmen 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

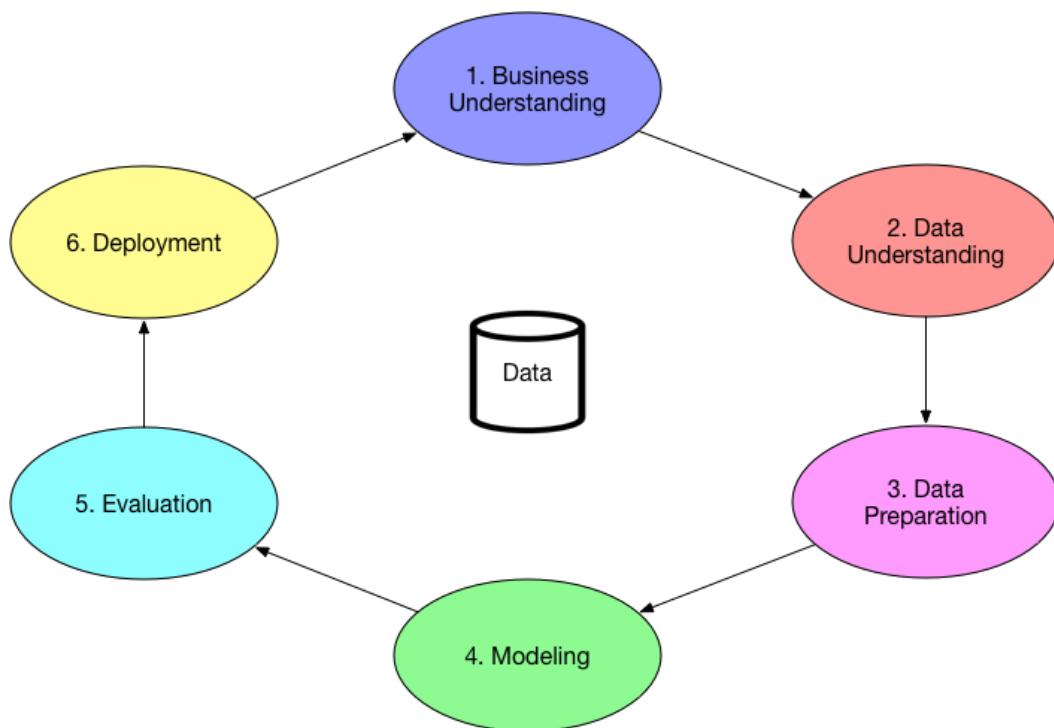


Abbildung 2.1: Konzeptionelles CRISP-DM Modell

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 Kassiovorgang 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 Datenbanksysteme 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

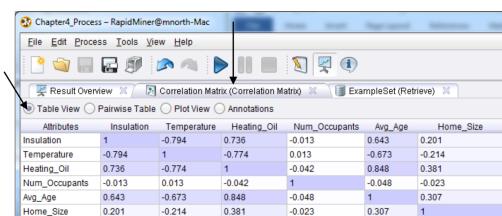
In diesem Schritt müssen die gesammelten Rohdaten verarbeitbar gemacht werden. Beispielsweise müssen verfälschte bzw. für die Analyse unwichtige Daten herausgefiltert, Attribute und deren Typen transformiert oder generell Datensätze bereinigt werden. Letztendlich müssen die Daten so aufbereitet werden, 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



The screenshot shows the RapidMiner interface with a correlation matrix table titled "Correlation Matrix (Correlation Matrix)". The table has columns labeled "Attributes", "Insulation", "Temperature", "Heating_Oil", "Num_Occupants", "Avg_Age", and "Home_Size". The rows follow the same pattern. The diagonal elements are all 1.0. The correlation values between attributes are as follows:

	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Insulation	1.0	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1.0	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1.0	-0.042	0.848	0.381
Num_Occupants	-0.013	0.013	-0.042	1.0	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1.0	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1.0

Abbildung 2.2: Ergebnis einer Korrelationsmatrix [North:2012]

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.

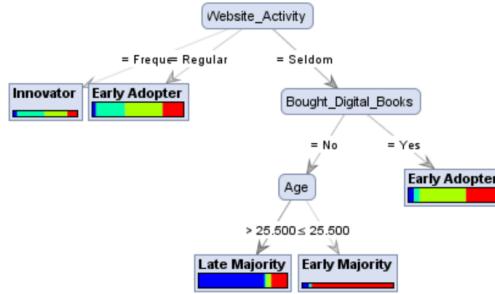


Abbildung 2.3: Ergebnis eines Entscheidungsbaums [North:2012]

Entscheidungsbäume sind eine grafische Darstellung von hierarchisch aufeinanderfolgenden Entscheidungen.

Prinzipiell versucht man, beim Modeling ein Modell zu erschaffen (bzw. zu „trainieren“), welches die echte Welt 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 falscher Parameter im Algorithmus oder durch ungenaue Daten, fehlerhaft sein. Ist dies der Fall, muss entweder das Modell entsprechend angepasst werden oder 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 kommt. Betrachtet man die Ergebnisse tauchen neben einigen spezifischen Prozessen auch der sog. SEMMA Prozess und der KDD Prozess auf.

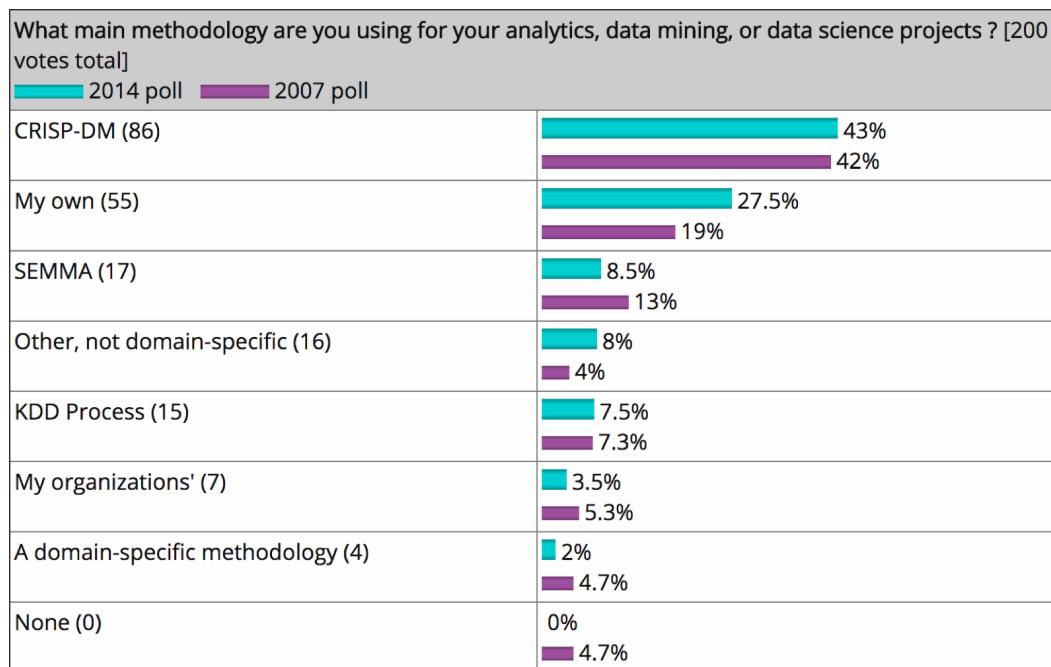


Abbildung 2.4: Umfrage welche DM-Prozesse im Unternehmen zum Einsatz kommen [Kdn]

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.

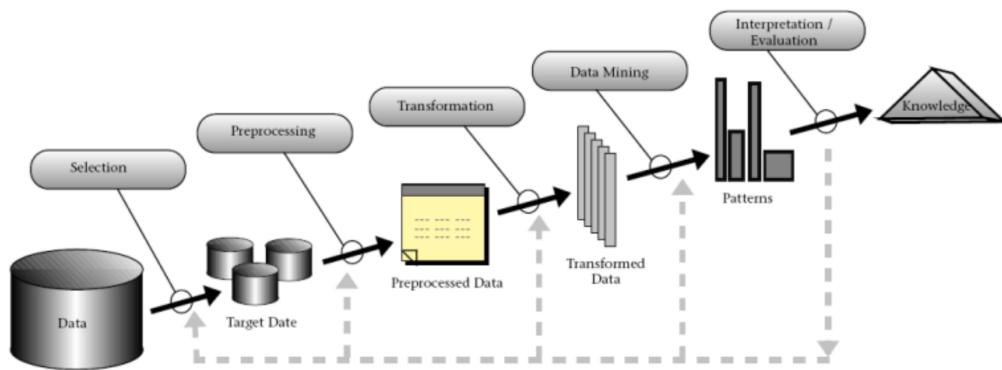


Abbildung 2.5: Schritte die den KDD Prozess zusammensetzen [M.96]

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 [Unk]:

- 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 Understanding“ (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.

Vorstellung der Software

3.1 Rapidminer Studio

Bei RapidMiner handelt es sich um eine „Open Source Data Science Platform“, welche vom gleichnamigen Unternehmen unter der AGPL 3.0 Lizenz vertrieben wird. Der Quellcode ist offen und kann auf Github gefunden werden. Neben dem im Folgenden näher untersuchten „RapidMiner Studio“ bietet die Plattform noch die Produkte „RapidMiner Server“ zur Automatisierung von Prozessen und „RapidMiner Radoop“ welches es erlaubt Berechnungen in einer Hadoop Umgebung auf einem Cluster auszuführen. Für kleinere Datensätze mit bis zu 10.000 Einträgen steht eine kostenlose „Education Version“ zur Verfügung. Außerdem bietet der Hersteller die Software für sämtliche Plattformen (Windows, MacOS, Linux) an.

Das Tool „RapidMiner Studio“ bietet die Möglichkeit einen Datenverarbeitungsprozess mithilfe einer eingängigen Oberfläche zu erstellen. Dafür muss die Software zunächst (auf einer der oben genannten Plattformen) installiert werden.

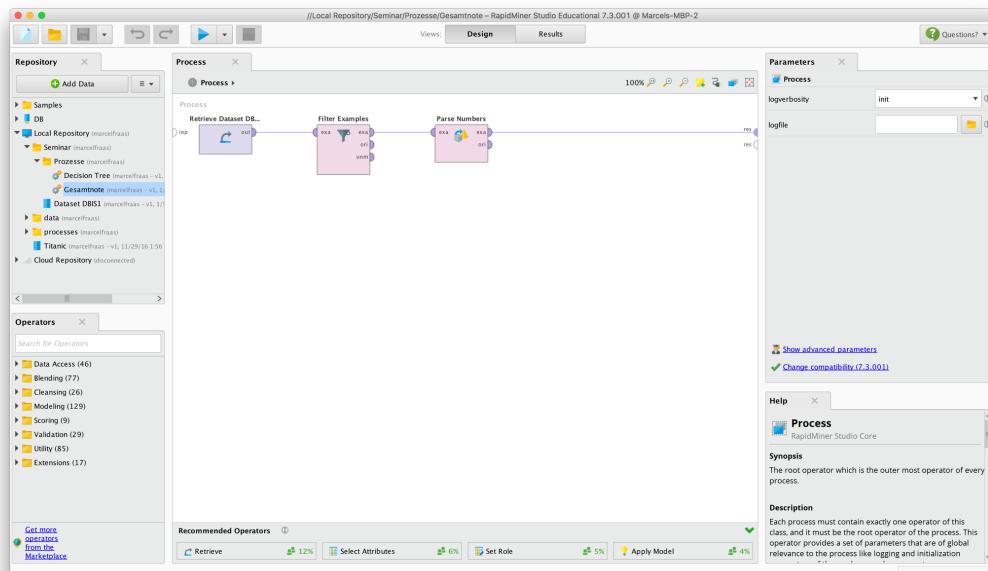


Abbildung 3.1: Die RapidMiner Design View

Die Oberfläche bietet ein Intuitives User Interface, welches die Einarbeitung sehr vereinfacht. Im Abschnitt links oben befindet sich das sog. „Repository“. Hier befinden sich die konfigurierten Datenquellen, welche bspw. in Form von Excel-, Access-,

SAS oder SPSS Dateien vorliegen können. Alternativ können auch direkt SQL Datenbanken, wie MySQL, Microsoft SQL Server, Oracle, PostgreSQL u.a., angebunden werden.

Darunter befindet sich die Ordnerstruktur der „Operators“. Von hier können die zahlreichen Operatoren per Drag and Drop in den Hauptbereich der Design-Ansicht, dem sog. „Process“ gezogen werden.

In der Prozessübersicht können Operatoren und Repositories angeordnet und miteinander verknüpft werden.

Auf der rechten Seite befinden sich, neben der Hilfe, noch eine Übersicht über die Parameter des aktuell ausgewählten Operators.

Zu guter Letzt bietet RapidMiner Studio noch eine zusätzliche Hilfsfunktion, die sog. „Recommended Operators“. Hier werden mit Hilfe von statistischen Auswertungen die am wahrscheinlichsten, zum aktuellen Prozess passenden, nächsten Operatoren angezeigt. Dies erleichtert den Einstieg in die Software ungemein.

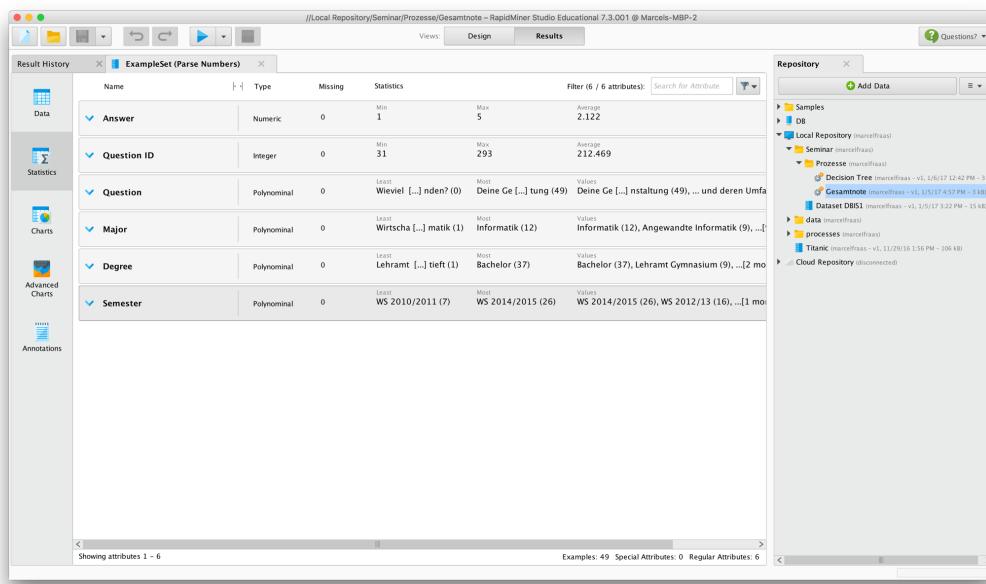


Abbildung 3.2: Die RapidMiner Result View

Führt man einen Prozess aus, bzw. wählt man manuell in der Menüleiste den Reiter „Results“ bekommt man das Ergebnis entweder in roh Form, also als Tabelle, in einer statistischen Übersicht oder als Graph bzw. Diagramm.

Generell ist die Einarbeitung in RapidMiner Studio, gerade auch wegen den detaillierten Hilfsfunktionen, sehr eingängig und führt schnell zu Ergebnissen.

3.2 Microsoft Azure Machine Learning Studio

Das „Microsoft Azure Machine Learning Studio“ ist, wie der Name bereits verrät, teil der Microsoft Cloud Umgebung Azure. Folglich ist das Tool eine Webanwendung, welche per Browser gestartet werden kann. Für den Einstieg benötigt man lediglich einen Microsoft Azure Account, welchen man kostenlos erstellen kann. Damit hat man dann vollen Zugriff auf alle Funktionen des Machine Learning Studios. Einschränkungen des kostenlosen Accounts gibt es hier nur bezüglich der Rechenzeit und den weiteren Services der Azure Umgebung (APIs etc.).

Neben dem Machine Learning Studio bietet die Microsoft Cloud eine Vielzahl an Services und Ressourcen, wie virtualisierte Server, Hosting von Webanwendungen und APIs sowie Datenbanken. Durch den „Platform as a Service“ Aspekt, wird eine extrem einfache Kollaboration und schnelles Deployment ohne eigenen Administrationsaufwand gewährleistet [Bar15].

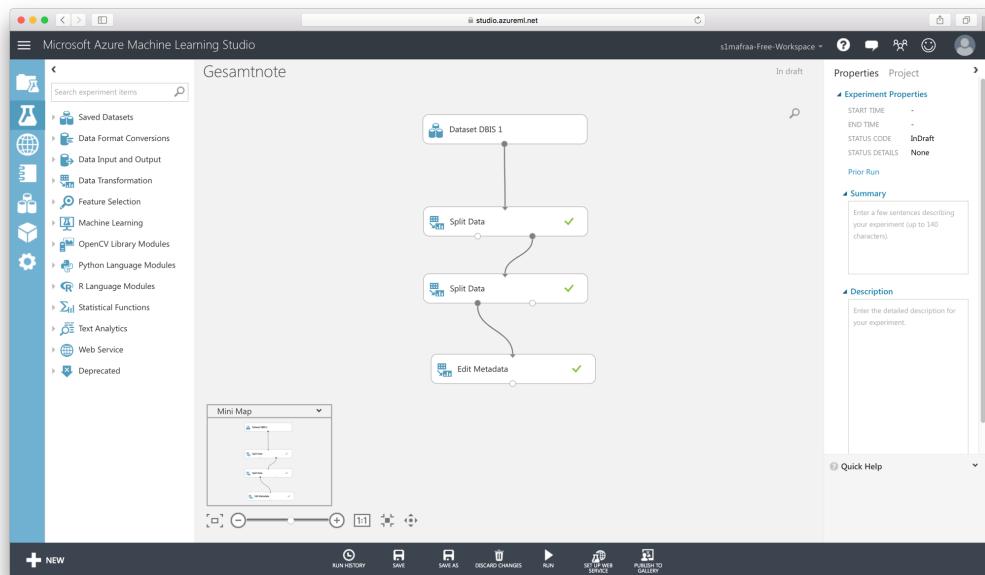


Abbildung 3.3: Ein Experiment im MSA Machine Learning Studio

Nach dem starten der Webanwendung kann der Anwender direkt damit beginnen ein neues Experiment anzulegen. Die Oberfläche ist ähnlich dem RapidMiner User Interface aufgebaut. Auf der linken Seite befinden sich die Datenquellen und die Operatoren. Unterstützt werden von Microsoft u.a. die Excel-, CSV-, ARFF-, sowie Plain Text und RObject Dateiformate. Außerdem können natürlich ebenfalls Daten aus einer SQL Datenbank, wie beispielsweise einer Microsoft SQL Serverinstanz aus der Azure Cloud, geladen werden.

Auch hier können die Operatoren per Drag and Drop in die Arbeitsfläche gezogen werden und mit den Parametern am rechten Rand konfiguriert werden.

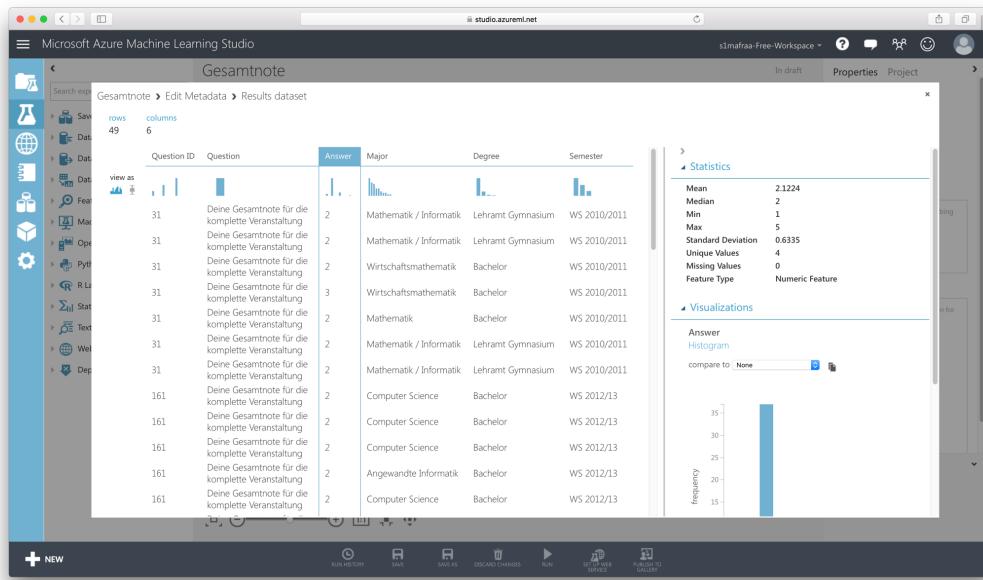


Abbildung 3.4: Das Ergebnis eines Experiments im MSA Machine Learning Studio

Die Visualisierung des Ergebnisses ist im Machine Learning Studio abhängig vom jeweiligen Experiment. In diesem Fall wird dieses als Tabelle mit einigen Statistiken und Diagrammen dargestellt.

Aufgrund der Tatsache, dass es sich beim Microsoft Azure Machine Learning Studio um eine Webanwendung handelt, müssen ein paar Abstriche in Sachen Bedienbarkeit gemacht werden. Dafür ist die Kollaboration mit Anderen und die Einbindung in Webservices erheblich vereinfacht, was besonders zum Tragen kommt, wenn man bereits eine gewisse Infrastruktur innerhalb der Azure Cloud besitzt.

Datensatz und Beispiel-Modell

4.1 Datensatz

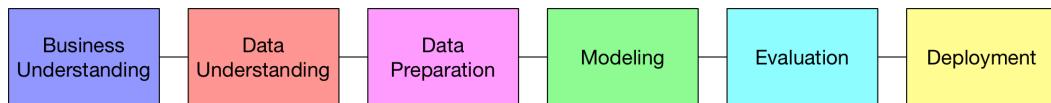


Abbildung 4.1: Der CRISP-DM Prozess als Kette der einzelnen Schritte

Für die Umsetzung des Beispiels wurde der Eingangs bereits beschriebene CRISP-DM Data Mining Prozess verwendet. Die Daten wurden gemäß den 6 Schritten „Business Understanding“, „Data Understanding“, „Data Preparation“, „Modeling“, „Evaluation“ und „Deployment“ verarbeitet.

4.1.1 Business Understanding

Als Datengrundlage für das Beispiel dienen die Daten der Veranstaltungsevaluationen der Vergangenen Semester, durchgeführt von der Fachschaft Mathematik, Physik und Informatik der Universität Bayreuth. Mithilfe der vorliegenden Datengrundlage sollen nun folgende Fragen beantwortet werden:

- „Warum beschweren sich einige Studenten, während andere die Veranstaltung super finden? Bzw. welche Ursachen gibt es für eine Beschwerde?“
- „Wie sind die Prognosen für die bevorstehenden Evaluationen?“

4.1.2 Data Understanding

Im nächsten Schritt ist es wichtig herauszufinden, wie die Daten erhoben wurden. Im Beispiel der Veranstaltungsevaluationen ist dies durch austeilen eines Fragebogens erfolgt.

Anhand des Fragebogens erkennt man, dass die Antwortmöglichkeiten in Form von anzukreuzenden Kästchen gegeben waren. Jedes Kreuz repräsentiert dabei einen Wert im Bereich von 0 bis max. 4.

Vorlesung:	Dozent (ganzer Name):
genauer Studiengang (mit evtl. Nebenfach):	Fachsemester:

1. Fragen zur Vorlesung

a) Ist der Inhalt der Vorlesung strukturiert?	sehr <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	überhaupt nicht <input type="checkbox"/>
b) Ist Tafelbild / die Präsentation strukturiert und verständlich?	sehr <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	überhaupt nicht <input type="checkbox"/>
c) Wie findest du die Geschwindigkeit der Vorlesung?	zu langsam <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	zu schnell <input type="checkbox"/>
d) Bist du mit der Nutzung zusätzlicher Medien zufrieden?	sehr <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	überhaupt nicht <input type="checkbox"/>
e) Geht der Dozent auf seine Hörer ein, regt er zu Fragen an?	häufig <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	selten <input type="checkbox"/>
f) Wie beantwortet der Dozent gestellte Fragen?	verständlich <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	unverständlich <input type="checkbox"/>
g) Vermag der Dozent den Inhalt zu motivieren?	ja <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	überhaupt nicht <input type="checkbox"/>
h) Falls es ein vorgefertigtes Skript gibt, wie hilfreich und ergänzend zur Vorlesung ist es?	sehr <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	überhaupt nicht <input type="checkbox"/>
i) Sollte der Dozent ein Skript zur Vorlesung anbieten?	□ ja <input type="checkbox"/>	□ nein <input type="checkbox"/>				

2. Fragen zu den Übungen:

a) Bewerte die Schwierigkeit der Übungsaufgaben...	unlösbar <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	zu leicht <input type="checkbox"/>
b) ...und deren Umfang!	zu viel <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	□ <input type="checkbox"/>	zu wenig <input type="checkbox"/>

Abbildung 4.2: Fragebogen, zur Erhebung der Evaluationsdaten

Außerdem muss festgestellt werden, wie die Daten in der Datenbank abgelegt werden. Hierfür kann man bspw. ein ER-Diagramm erzeugen oder sich eine Vorschau der Daten in Form einer Tabelle generieren lassen.

The screenshot shows two MySQL database tables:

- survey_returned_values** (left):
 - Fields: id (INT(11)), returned_questionnaire_id (INT(11)), question_id (INT(11)), survey_id (INT(11)), value_int (INT(11)), value_text (TEXT).
 - Indexes: None.
- survey_returned_questionnaires** (right):
 - Fields: id (INT(11)), survey_id (INT(11)), trainer_id (INT(11)), course_id (INT(11)), semester (INT(11)).
 - Indexes: None.

Evaluation.csv

Question ID	Question	Answer	Major	Degree	Semester
1	Der Dozent schreibt auf...	1	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
2	Welche Medien verwendet der Dozent?	10	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
5	Wie würdest du das Tafelbild beurteilen?	2	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
6	Wie findest du die Geschwindigkeit der Vorlesung?	1	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
4	Ist die Vorlesung strukturiert?	1	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
7	Geht der Dozent auf seine Hörer ein, regt er zu Fragen an?	2	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
8	Vermag der Dozent den Inhalt zu motivieren?	2	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
10	Sollte der Dozent ein Skript zur Vorlesung anbieten (falls nicht bereits schon vorhanden?)	1	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
12	Bewerte die Schwierigkeit der Übungsaufgaben...	3	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
13	... und deren Umfang!	2	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011
14	Wie bewertest du deinen Übungsleiter?	0	Mathematik / Informatik	Lehramt Gymnasium	WS 2010/2011

Abbildung 4.3: Repräsentation der Daten in der Datenbank

Mithilfe dieser Informationen erkennt man, dass sich die gesuchten Daten in den beiden Tabellen „survey_returned_values“ und „survey_returned_questionnaires“ befinden und wie diese strukturiert sind.

4.1.3 Data Preparation

Wurden nun die entsprechenden Tabellen bzw. deren Attribute festgelegt müssen diese nun im nächsten Schritt aus der Datenbank extrahiert werden um separat analysiert werden zu können. Dies geschieht mit der folgenden SQL-Abfrage.

```
1 SELECT survey_questions.id AS 'Question ID',
2       survey_questions.description AS 'Question',
3       survey_returned_values.value_int AS 'Answer',
4       survey_course.name AS 'Major',
5       survey_course.degree AS 'Degree',
6       survey_terms.name AS 'Semester'
7   FROM `survey_returned_questionnaires`
8   INNER JOIN survey_course
9     ON survey_returned_questionnaires.course_id = survey_course.id
10  INNER JOIN survey_surveys
11    ON survey_returned_questionnaires.survey_id = survey_surveys.id
12  INNER JOIN survey_terms
13    ON survey_surveys.term_id = survey_terms.id
14  INNER JOIN survey_returned_values
15    ON survey_returned_questionnaires.id = survey_returned_values.returned_questionnaire_id
16  INNER JOIN survey_questions
17    ON survey_returned_values.question_id = survey_questions.id
18 WHERE survey_returned_questionnaires.survey_id=23
19   OR survey_returned_questionnaires.survey_id=157
20   OR survey_returned_questionnaires.survey_id=251
```

Abbildung 4.4: SQL Anfrage für den Export der Daten

Die Ergebnisdatensätze der SQL-Abfrage können nun als .csv Datei exportiert werden um danach in der jeweiligen Data Mining Software verarbeitet werden zu können.

4.1.4 Modeling

In der Modellierungsphase muss nun ein Algorithmus gewählt werden, welcher die vorbereiteten Daten so verarbeiten kann, dass die Anfangs definierten Fragestellungen beantwortet werden können. In unserem Beispiel eignet sich die Frage nach der Gesamtnote, da dieses Attribut eine gute Repräsentation der Gesamtbewertung für eine Veranstaltung darstellt.

Es muss also ein Algorithmus gewählt werden, welcher nach einem einzelnen Attribut, dem sog. Label, Daten Mined. Ein einfach auszuwertender Algorithmus, der genau diese Voraussetzungen erfüllt ist der sog. „Entscheidungsbaum“ (engl.: „Decision Tree“).

Die Knoten des Baumes stellen klassifizierende Attribute des Ausgangs-Datensatzes dar, während die Blätter die möglichen Werte des Labels anzeigen. Außerdem zeigen die Blätter noch an, auf wieviel Datensätzen die jeweilige Klasse beruht.

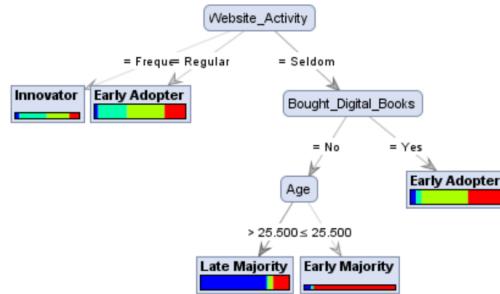


Abbildung 4.5: Beispiel für einen Entscheidungsbaum [North:2012]

4.1.5 Evaluation + Deployment

Schließlich kann mithilfe des „Decision Trees“ ermittelt werden, dass beispielsweise das Studienfach oder ein bestimmtes Semester die Ursache für vermehrte Beschwerden (respektive mehrere schlechte Bewertungen) waren. Gleichzeitig kann man mit den entsprechenden Informationen über Studienfach, Abschluss etc. der aktuellen Kursteilnehmer eine Prognose für die bevorstehende Lehrveranstaltungsevaluation abgeben.

4.2 Umsetzung

4.2.1 Rapidminer Studio

Um das Beispieldaten in RapidMiner Studio umzusetzen wurde der folgende Prozess modelliert:

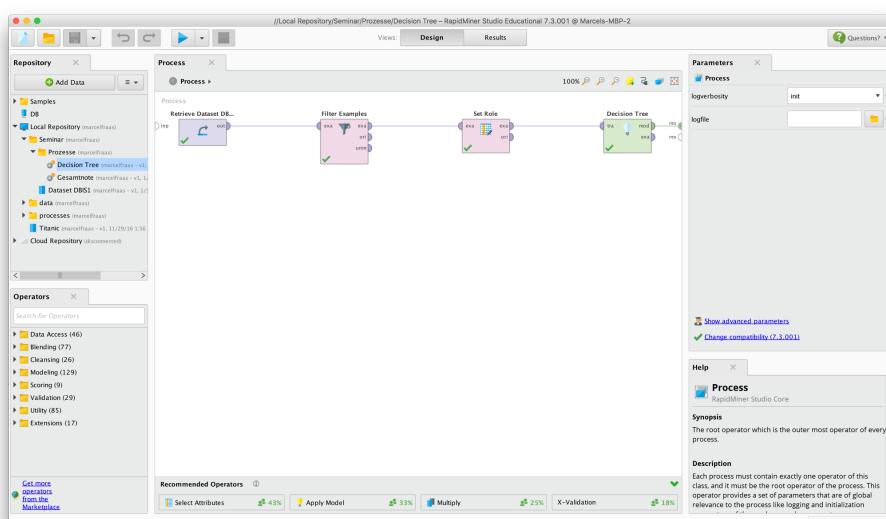


Abbildung 4.6: Der Beispielprozess im RapidMiner Studio

Zunächst wurde die .csv Datei mit dem „Retrieve Dataset“ Operator in den Prozess geladen. Anschließend wurden per „Filter Example“ sowohl alle Datensätze mit NULL Einträgen entfernt, als auch die Einträge, welche die Frage nach der Gesamtnote beinhalten, extrahiert. Danach wurde die Spalte „Answer“, welche den Wert der Note beinhaltet als „Label“ definiert. Zuletzt wurden die Daten dann dem „Decision Tree“ Algorithmus übergeben und das Ergebnis zurück in die Anzeige geladen.

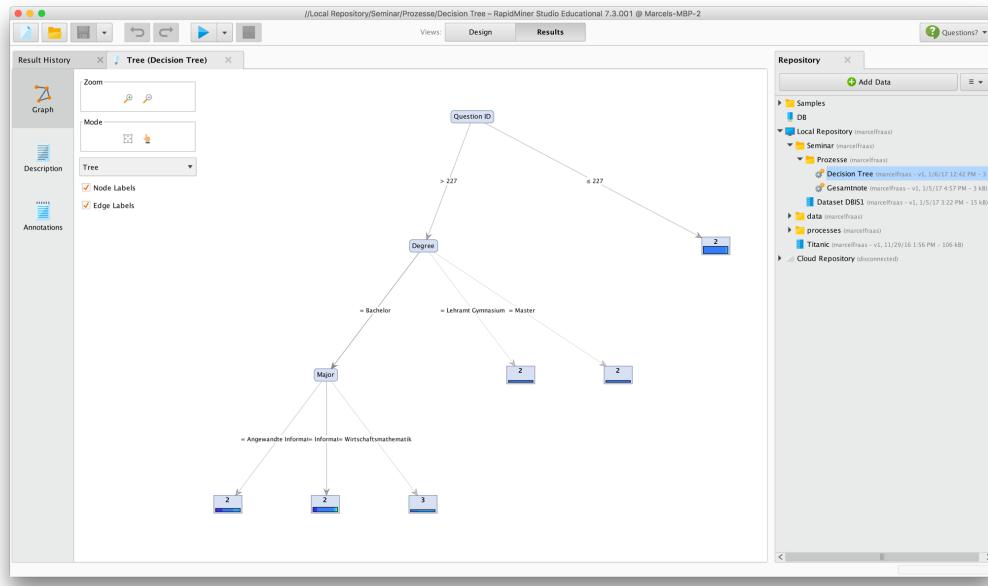


Abbildung 4.7: Das Ergebnis im RapidMiner Studio

Das Ergebnis ist ein Baum, an welchem erkennbar ist, dass in den meisten Fällen die Note „2“ für die Lehrveranstaltung vergeben wurde. Einzig bei der Betrachtung des Studienfaches fällt auf, dass Studierende, welche dem „Major“ Wirtschaftsmathematik angehören, eine 3 vergeben haben. An den Blättern lässt sich außerdem ablesen, wie groß die sog. „Confidence“ ist, also auf wie vielen Daten, die jeweilige Entscheidung basiert. Kenntlich gemacht wird dies durch die jeweiligen farblichen Balken. Diese sind wiefolgt zu interpretieren: Je dicker (höher) der Balken, desto mehr Datensätze liegen der Entscheidung zugrunde. Anhand (der Anzahl) der Farben erkennt man wieviel unterschiedliche Werte für die jeweilige Klasse gefunden wurden und entsprechend an der „breite“ der Farbe kann man ablesen wie oft der jeweilige Wert in Relation aufgetaucht ist.

4.2.2 Microsoft Azure Machine Learning Studio

Das korrespondierende Experiment im Microsoft Azure Machine Learning Studio sieht folgendermaßen aus:

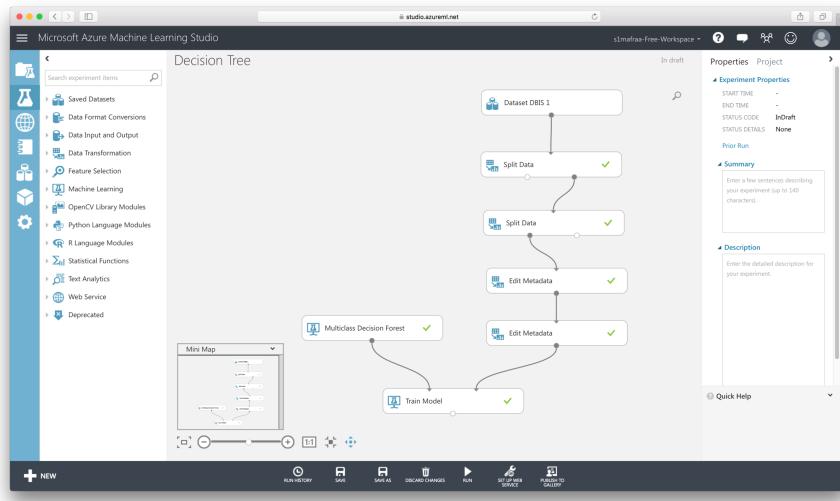


Abbildung 4.8: Das Beispiel-Experiment im Microsoft Azure Machine Learning Studio

Hier fällt auf der rechten Seite auf, dass je 2 „Split Data“ und „Edit Metadata“ Operatoren gebraucht wurden. Um die selbe Filter-Funktionalität wie beim RapidMiner Studio Prozess abzubilden, mussten zunächst alle NULL Einträge und anschließend ebenfalls alle Datensätze, welche die Frage nach der Gesamtnote beinhalten, herausgefiltert werden. Dafür waren hier 2 separate „Split“ Operationen notwendig. Des Weiteren wurde sowohl für die Definition des Attributs „Answer“ als Label, als auch für die Definition der restlichen Attribute als „klassifizierende Attribute“ je ein „Edit Metadata“ Operator benötigt. Zuletzt fällt noch auf, dass anstatt eines Decision Tree Algorithmus ein „Multiclass Decision Forest“, welcher als Konfiguration eine Decision-Tree Anzahl von 1 bekommen hat, zur Berechnung verwendet wurde. Dies war notwendig um überhaupt ein auswertbares Ergebnis zu erhalten, welches zumindest Ansatzweise mit dem Ergebnis des RapidMiner Studios vergleichbar ist.

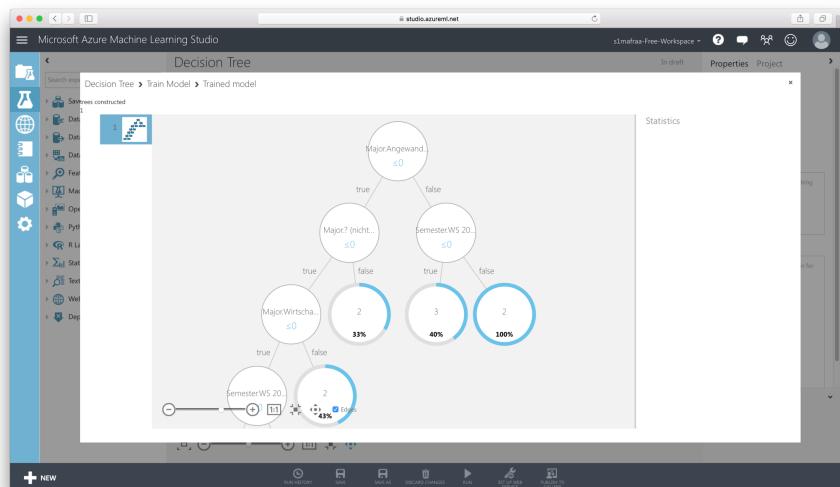


Abbildung 4.9: Das Ergebnis im Microsoft Azure Machine Learning Studio

Wie aufgrund der Tatsache, dass ein anderer Algorithmus verwendet wurde, anzunehmen ist, weicht das Ergebnis vom vorherigen Ergebnis des RapidMiner Studios in einigen Punkten ab. Der Baum zeigt, dass nicht zwangsläufig das Studienfach ausschlaggebend für eine schlechtere Bewertung war, sondern das Semester in welchem die Umfrage durchgeführt wurde. Auch hier lässt sich anhand der Blätter ablesen, wie groß die „Confidence“ der jeweiligen Klasse ist. So lässt sich also interpretieren, dass Studierende, die nicht dem „Major“ Angewandte Informatik angehören, dafür allerdings im Wintersemester 2012/13 an der Umfrage teilgenommen haben überwiegend die Note „3“ vergeben haben. Die Prozentzahl zeigt dabei an, dass 40% der zugrunde liegenden Datensätze in die Klasse „3“ fallen, wohingegen die anderen 60% sich auf die restlichen Werte 1, 2 bzw. 4, 5 und 6 verteilen.

Fazit

Zum Abschluss werden beide Tools noch einmal in Bezug auf unterschiedliche Aspekte miteinander verglichen.

1. Installation und Setup:

Zunächst konnte das „Setup“ bei beiden Tools ohne Probleme durchgeführt werden. Allerdings besitzt das Microsoft Azure Machine Learning Studio hier den Vorteil, dass keine Installation (sondern nur eine Registrierung bzw. ein Login) notwendig ist und man direkt mit der Analyse seiner Daten beginnen kann. Das RapidMiner Studio ist zwar für alle Plattformen erhältlich, muss aber dennoch (manuell) installiert werden. Zudem wird eine aktuelle Java Runtime vorausgesetzt, welche u.U. auch erst installiert werden muss. Microsoft setzt diesbezüglich nur einen (aktuelleren) Internetbrowser voraus.

2. Features und Funktionen:

Beide Programme kommen mit einer Vielzahl an Implementierungen verschiedener Algorithmen. Bezüglich der Anzahl hat das RapidMiner Studio jedoch etwas mehr zu bieten.

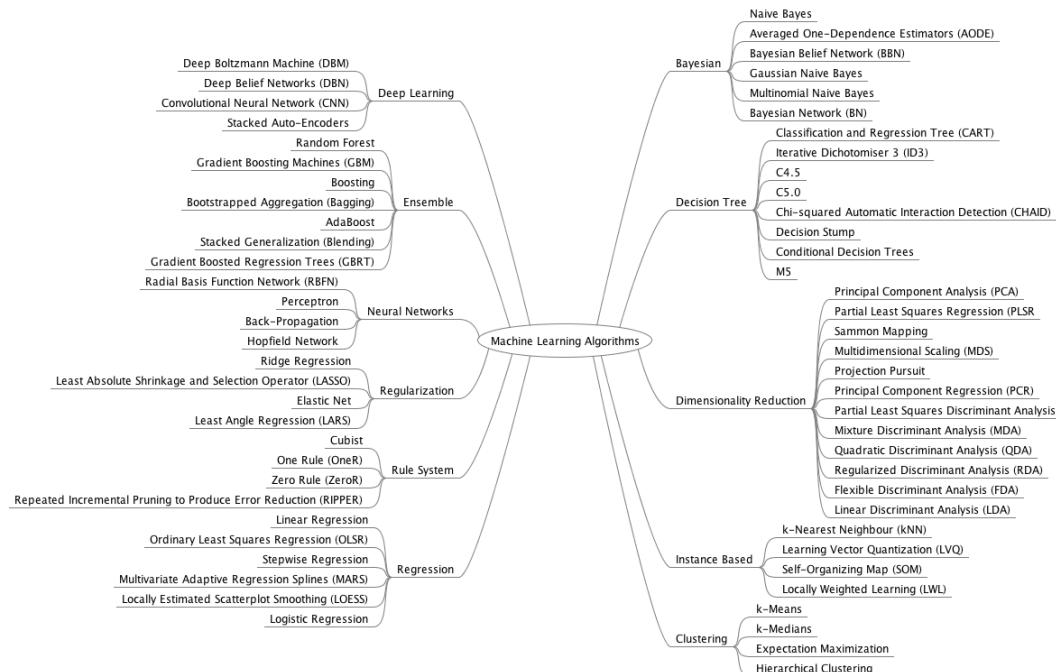


Abbildung 5.1: Algorithmen des RapidMiner Studios [Mlm]

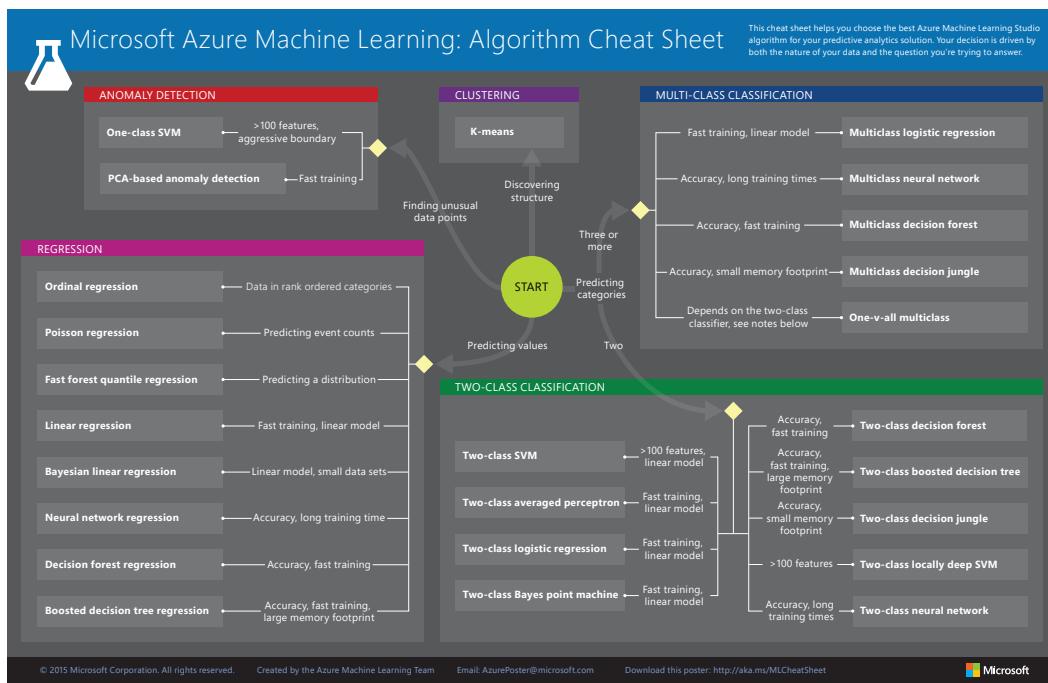


Abbildung 5.2: Algorithmen des Microsoft Azure Machine Learning Studios [Msa]

3. Kollaboration und Zusammenarbeit:

Gleicher wie bei der Automatisierung gilt auch für die Unterstützung eines kollaborativen Workflows mit mehreren Anwendern. Während in Microsofts Machine Learning Studio die Experimente direkt mit anderen Anwendern geteilt bzw. zusammen an diesen gearbeitet werden kann, benötigt es auf der RapidMiner Plattform eines eigenen Servertools (sofern die Prozesse nicht als Datei abgespeichert und per Email verschickt werden wollen).

4. Bedienung und User Interface:

Bezüglich der Bedienbarkeit bietet das RapidMiner Studio einen angenehm hohen Komfort, während das Machine Learning Studio von Microsoft an die Webtechnologien und Möglichkeiten eines Internetbrowsers gebunden ist. Nichts desto trotz ist die nötige Einarbeitungszeit bei beiden Tools sehr kurz.

5. Schnittstellen und Erweiterungsmöglichkeiten:

Da das RapidMiner Studio Quelloffen ist, ist es hier möglich durch Plugins die Feature-Liste selbst noch zu erweitern und eigene Algorithmen zu implementieren. Jedoch bietet das Microsoft Azure Machine Learning Studio hier einen deutlich höheren Komfort, indem es direkt das Einbinden von R Skripten, sowie Python Code ermöglicht.

6. Automatisierung und Deployment:

Die Experimente können aus dem Machine Learning Studio direkt als Webser-

vices exportiert und per REST-API angesteuert werden. Dies ermöglicht einen sehr effizienten und schnellen Workflow von der Analyse der Daten hin zur Implementierung eines automatisierten Modells in einer bestehenden Anwendung. RapidMiner Studio benötigt dagegen für das Deployment einen eigenen Serverinstallation auf Basis des RapidMiner Servers oder des RapidMiner Radoop Tools. Der Administrationsaufwand ist hier daher deutlich höher.

Abschließend lässt sich sagen, dass RapidMiner zurecht die populärere der beiden Plattformen ist, da diese intuitiver zu bedienen ist und in Sachen vorgefertigter Algorithmen deutlich mächtiger ist als die Lösung von Microsoft. Nichts desto trotz sollte auch, gerade wenn bereits ein Informationssystem auf Basis von Microsoft bzw. der Azure Cloud besteht, das Machine Learning Studio nicht außer Acht gelassen werden.

Literatur

- [Bar15] Jeff Barnes. *Azure Machine Learning*. Microsoft Press, 2015 (siehe S. 13).
- [M.96] Fayyad U. M. *Advances 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. 3).
- [Unk] Unknown. „KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW“. In: *Repositório Científico do Instituto Politécnico do Porto* () (siehe S. 9).

Websites

- [Duh12] Charles Duhigg. *How Companies Learn Your Secrets*. 2012. URL: <http://www.nytimes.com> (siehe S. 1).
- [Kdn] *What main methodology are you using for your analytics, data mining, or data science projects ?* 2014. URL: <http://www.kdnuggets.com/> (siehe S. 8).
- [Mlm] *RapidMiner Machine Learning Algorithms*. URL: <http://machinelearningmastery.com> (siehe S. 23).
- [Msa] *Microsoft Azure Machine Learning: Algorithm Cheat Sheet*. URL: <https://docs.microsoft.com/> (siehe S. 24).

Abbildungsverzeichnis

2.1	Konzeptionelles CRISP-DM Modell	4
2.2	Ergebnis einer Korrelationsmatrix [North:2012]	6
2.3	Ergebnis eines Entscheidungsbaums [North:2012]	7
2.4	Umfrage welche DM-Prozesse im Unternehmen zum Einsatz kommen [Kdn]	8
2.5	Schritte die den KDD Prozess zusammensetzen [M.96]	9
3.1	Die RapidMiner Design View	11
3.2	Die RapidMiner Result View	12
3.3	Ein Expriment im MSA Machine Learning Studio	13
3.4	Das Ergebnis eines Expriments im MSA Machine Learning Studio	14
4.1	Der CRISP-DM Prozess als Kette der einzelnen Schritte	15
4.2	Fragebogen, zur Erhebung der Evaluationsdaten	16
4.3	Repräsentation der Daten in der Datenbank	16
4.4	SQL Anfrage für den Export der Daten	17
4.5	Beispiel für einen Entscheidungsbaum [North:2012]	18
4.6	Der Beispielprozess im RapidMiner Studio	18
4.7	Das Ergebnis im RapidMiner Studio	19
4.8	Das Beispiel-Experiment im Microsoft Azure Machine Learning Studio	20
4.9	Das Ergebnis im Microsoft Azure Machine Learning Studio	20
5.1	Algorithmen des RapidMiner Studios [Mlm]	23
5.2	Algorithmen des Microsoft Azure Machine Learning Studios [Msa]	24

