HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Big Data / Data Mining

Einführung

Christoph Menzer
Fachgruppe Informatik

26. Oktober 2020

Inhalt

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Motivation

Warum Data Mining?

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- in jeder Sekunde werden Tausende oder Millionen von Daten gemessen, erhoben und gespeichert
- Warum? Weil wir es können und es kaum oder kein Geld kostet
- sowohl im beruflichen als auch im privaten Umfeld
- Daten und (vor allem) daraus abgeleitete Information sind Handelsware und zu Produktionsfaktor geworden

Motivation

Was ist Data Mining?

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Data Mining ist der Prozess des Entdeckens bedeutsamer neuer Zusammenhänge, Muster und Trends durch die Analyse großer Datensätze mittels Mustererkennung sowie statistischer und mathematischer Verfahren. (Ittner 2019)

Data Mining [ist] eine Sammlung von Techniken, Methoden und Algorithmen für die Analyse von Daten, die somit auch Grundtechniken für neuere und komplexere Ansätze, wie das Business Intelligence oder auch Big Data darstellen. (Cleve und Lämmel 2014, S. 3)

Motivation

Was können Sie erwarten?

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- Techniken und Methoden zur Auswertung großer Datenmengen
- Einführung in Prozesse und Techniken der Mustererkennung in *strukturierten Daten* (Teilgebiet Data Mining, vgl. Text Mining, Web Mining, Künstliche Intelligenz (KI))
- Grundbegriffe und Vorgehensweisen des Data Mining
- Anwendungsklassen und Einsatzmöglichkeiten
- Verfahren zur Datenanalyse
- Datenselektion und -integration, Datensäuberung, Datenreduktion, Datentransformation (Datenvorverarbeitung, DVV)
- Bewertung und Visualisierung der Resultate

Überblick

Maschinelles Lernen

- computerbasierte Lernverfahren, die aus Eingabeinformationen Wissen (vgl. S. 15) generieren können
- einfachste Lernstrategie: Auswendiglernen (abspeichern, z.B. in Datenbank; hardcoded source code)
- angestrebt wird aber: Verständnis von Zusammenhängen und Hintergründen (Muster oder Abhängigkeiten)
- Induktives Lernen: aus Beispielen lernen und verallgemeinern
- statistische Verfahren zur Aufdeckung von Zusammenhängen

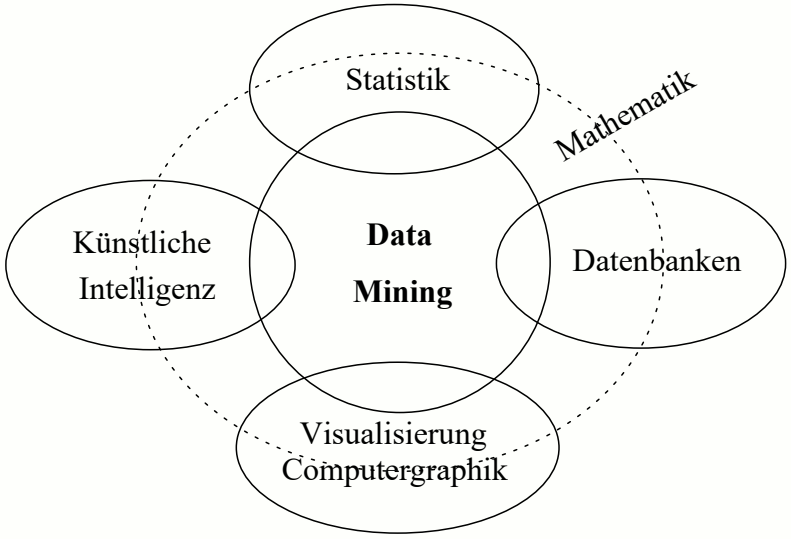
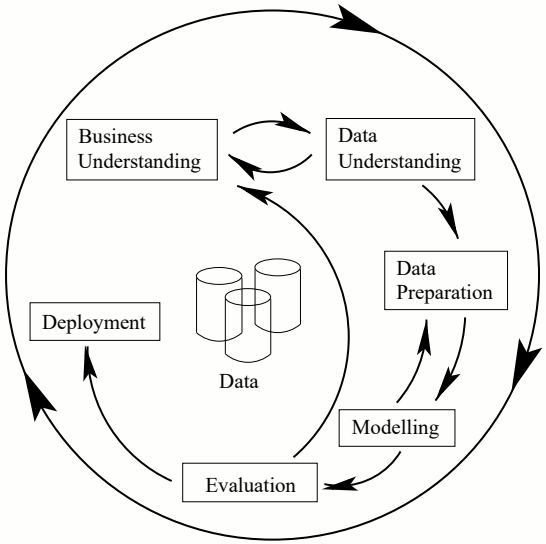


Abbildung: Interdisziplinarität (Quelle: Cleve und Lämmel 2014, S. 12)

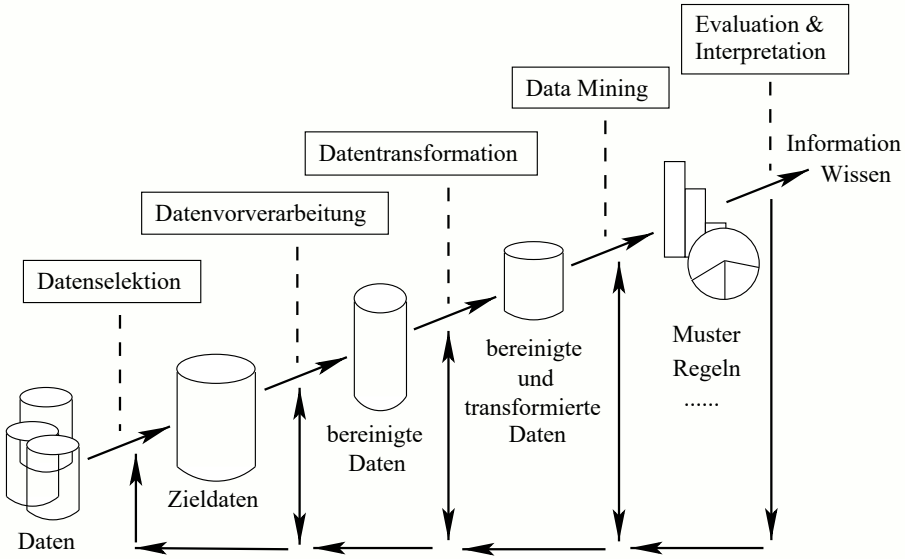
Überblick

Vorgehensmodelle

„Data Mining ist ein typisch iterativer Prozess: Man probiert etwas aus, berechnet ein Resultat und prüft dieses“ (Cleve und Lämmel 2014, S. 197)



(a) CRISP-Modell



(b) Ablauf nach Fayyad

Abbildung: Quelle: Cleve und Lämmel 2014, S. 5 f.

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

CRISP: Lebenszyklus in 6 Etappen

1. Verstehen der Aufgabe (Business understanding)
2. Verständnis der Daten (Data understanding)
3. Datenvorbereitung (Data preparation)
4. Modellbildung (Modeling)
5. Evaluation (Evaluation)
6. Einsatz im und Konsequenzen für das Unternehmen (Deployment)

Überblick

Anwendungsklassen

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Prognosemethoden

- Vorhersage numerischer Werte
- Klassifikation (Vorhersage)

Entdeckungsmethoden

- Cluster-Analyse (Zerlegung)
- Assoziationsanalyse (Zusammenhänge)

Überblick

Lernstrategien

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Unüberwachtes Lernen

- zu entdeckende Muster sind unbekannt
- Gruppierung oder Klassifikation ist nicht gegeben
- nicht vergleichbar
- z.B. Clustering

Überwachtes Lernen

- Resultat ist für Beispiele gegeben
- vergleichbar
- z.B. Vorhersage numerischer Werte, Klassifikation
- Annahme: Beispieldaten sind repräsentativ

Überblick

Analysetechniken

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- Lineare Regression
- k-Nearest Neighbour
- Bayes-Klassifikator
- Entscheidungsbaumlernen
- Support Vector Machine
- Künstliches Neuronales Netz
- k-Means
- A-Priori-Algorithmus

Überblick

Datenvorverarbeitung

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- schwierigste Aufgabe
- Ausreißer (fehlende, falsche oder widersprüchliche Werte)
- Datentransformation, Skalierung
- Qualität der Daten entscheidend für Ergebnis
- etwa 80% des Gesamtaufwandes
- Phase „Data understanding“ von großer Bedeutung
- Im ersten Schritt einfache statistische Tests durchführen
- Arten der Datenvorbereitung
  - Datenselektion und -integration
  - Datensäuberung
  - Datenreduktion
  - Datentransformation

Überblick

Werkzeuge

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- Texteditoren
- Tabellenkalkulationsprogramme
- ...
- KNIME (Konstanz Information Miner)
- WEKA (Waikato Environment for Knowledge Analysis)
- JavaNNS für SNNS (Stuttgarter Neuronale Netze Simulator)
- RapidMiner
- IBM SPSS Modeler
- R
- ...

In vielen Data-Warehouse- und Datenbank-Systemen sind Data-Mining-Komponenten integriert.

Überblick

Werkzeuge

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- Texteditoren
- Tabellenkalkulationsprogramme
- ...
- KNIME (Konstanz Information Miner)
- WEKA (Waikato Environment for Knowledge Analysis)
- JavaNNS für SNNS (Stuttgarter Neuronale Netze Simulator)
- RapidMiner
- IBM SPSS Modeler
- R
- ...

In vielen Data-Warehouse- und Datenbank-Systemen sind Data-Mining-Komponenten integriert.

Grundlagen

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Definition

Mit *Daten* bezeichnet man eine Ansammlung von Zeichen mit der dazugehörigen Syntax.

- Plural des Wortes Datum
- Interpretation als Informationseinheit
- Man unterschiedet
  - Unstrukturierte Daten
  - Semistrukturierte Daten
  - Strukturierte Daten

Aus Daten entstehen Informationen dadurch, dass diese Daten eine Bedeutung bekommen.

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Definition
Eine *Information* ist ein Datum, welches mit einer Bedeutung gekoppelt ist.

- zweckbestimmte Interpretation von Daten
- Daten werden erst dann zur Information, wenn sie im Kontext betrachtet werden und eine Bedeutung erhalten

Definition
Eine Information in Verbindung mit der Fähigkeit, diese zu benutzen, wird als *Wissen* bezeichnet

- eine Information wird zu Wissen, wenn man mit ihr etwas anzufangen weiß

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Definition

Data Mining (Datenschürfen) ist die Extraktion von Wissen aus Daten.

- Extraktion von implizitem Wissen
- dieses sollte nicht trivial, sondern nützlich sein
- weitgehend automatischer Ablauf (Vorbereitung der Daten aber oft manuell)
- Teilgebiete
  - Text Mining (unstrukturierte Daten)
  - Web Mining (semistrukturierte Daten)
  - Data Mining im engeren Sinn (strukturierte Daten)

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Datensatz, *Instanz*, *Objekt* oder *Muster*

- im Data Mining weitgehend synonym
- Damit ist stets ein Datensatz in einer Datenbanktabelle gemeint, der ein Objekt oder eine Instanz anhand einer Menge von Merkmalswerten charakterisiert

Wetter-Beispiel

```
(sunny, hot, high, false, no)
```

= math. Quintupel

- Auffassung als Objekt
- dieses Objekt ist eine Instanz der Klasse aller Wetter-Situationen

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

- Metrische
  - berechenbar (z.B. Mittelwerte)
  - miteinander vergleichbar
  - sortierbar (Ordnungsrelation, z.B. „<“ oder „>“)
- Ordinale
  - nicht berechenbar, aber sortierbar
- Nominale
  - nicht berechenbar
  - unterliegen keinerlei Rangfolge, d.h. sie lassen sich nicht sortieren

Metrische Datentypen können weiter unterteilt werden:

- Diskrete
  - endliche Anzahl von Werten, z.B. Schulnoten
- Kontinuierliche
  - beliebiger Zahlenwert innerhalb des Definitionsbereiches
  - im Computer auch diskret, beliebige Genauigkeit angenommen

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Weiter wird nach der zugrundeliegenden Skala unterschieden:

- Intervallbasierte
  - Nullpunkt willkürlich
  - eingeschränkt berechenbar (20°C nicht viermal wärmer als 5°C)
- Verhältnisbasierte
  - natürlicher Nullpunkt
  - berechenbar, vergleichbar (60km doppelt so weit wie 30km)
  - abhängig von Maßeinheit (60km ungleich 60cm)
- Absolutskalenbasierte
  - wie Verhältnisbasierte, aber unabhängig von Maßeinheit

Achtung

Numerische Werte sind nicht dringend metrisch (z.B. Kundennummer, PLZ).

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Beispiele

- metrisch, Intervallskala
  - Jahreszahlen
  - Temperatur in Celsius, Fahrenheit
- metrisch, Verhältnisskala
  - Entfernung
  - Körpergröße
- metrisch, Absolutskala
  - Lebensjahre
  - Zahl der Kinder

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Beispiele

- ordinal
  - Schulnoten {1,2,3,4,5,6}
  - Meinung {stimme zu, stimme teilweise zu, stimme nicht zu}
- nominal
  - Geschlecht {weiblich, männlich}
  - Haarfarbe {blond, dunkelblond, braun, rot, schwarz, grau}

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

- im Data Mining werden Datensätze oft miteinander verglichen
- Ähnlichkeit zweier Datensätze muss quantifizierbar sein
- meist über sogenannte Abstandsmaße realisiert, die die „Unähnlichkeit“ der Datensätze quantifizieren

Definition

Der *Abstand* zweier Datensätze v und w ist durch das *Abstandsmaß*

$$\text{dist}(v, w)$$

definiert.

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Definition

Die *Ähnlichkeit* zweier Datensätze v und w ist in Abhängigkeit von einem Abstandsmaß dist definiert.

$$\text{simil}(v, w) = f(\text{dist}(v, w))$$

- Je größer die Distanz zwischen den Datensätzen, desto geringer ist die Ähnlichkeit

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Eine Abstandsfunktion (auch *Distanzfunktion* genannt) muss folgende Eigenschaften erfüllen:

1. $\text{dist}(x, y) \geq 0$
2. $\text{dist}(x, x) = 0$ bzw. $\text{dist}(x, y) = 0$ gdw. $x = y$
3. $\text{dist}(x, y) = \text{dist}(y, x)$ (kommutativ)
4. $\text{dist}(x, y) \leq \text{dist}(x, z) + \text{dist}(z, y)$ (transitiv)

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Typische Distanzfunktionen:

Hamming-Distanz $\text{dist}_H(v, w) = \text{count}_i(v_i \neq w_i)$

Euklidische Distanz $\text{dist}_E(v, w) = \sqrt{\sum_i (v_i - w_i)^2}$

Manhatten-Distanz $\text{dist}_{Man}(v, w) = \sum_i |v_i - w_i|$

Maximum-Distanz $\text{dist}_{Max}(v, w) = \max(|v_i - w_i|)$

RapidMiner

RapidMiner

RapidMiner Studio

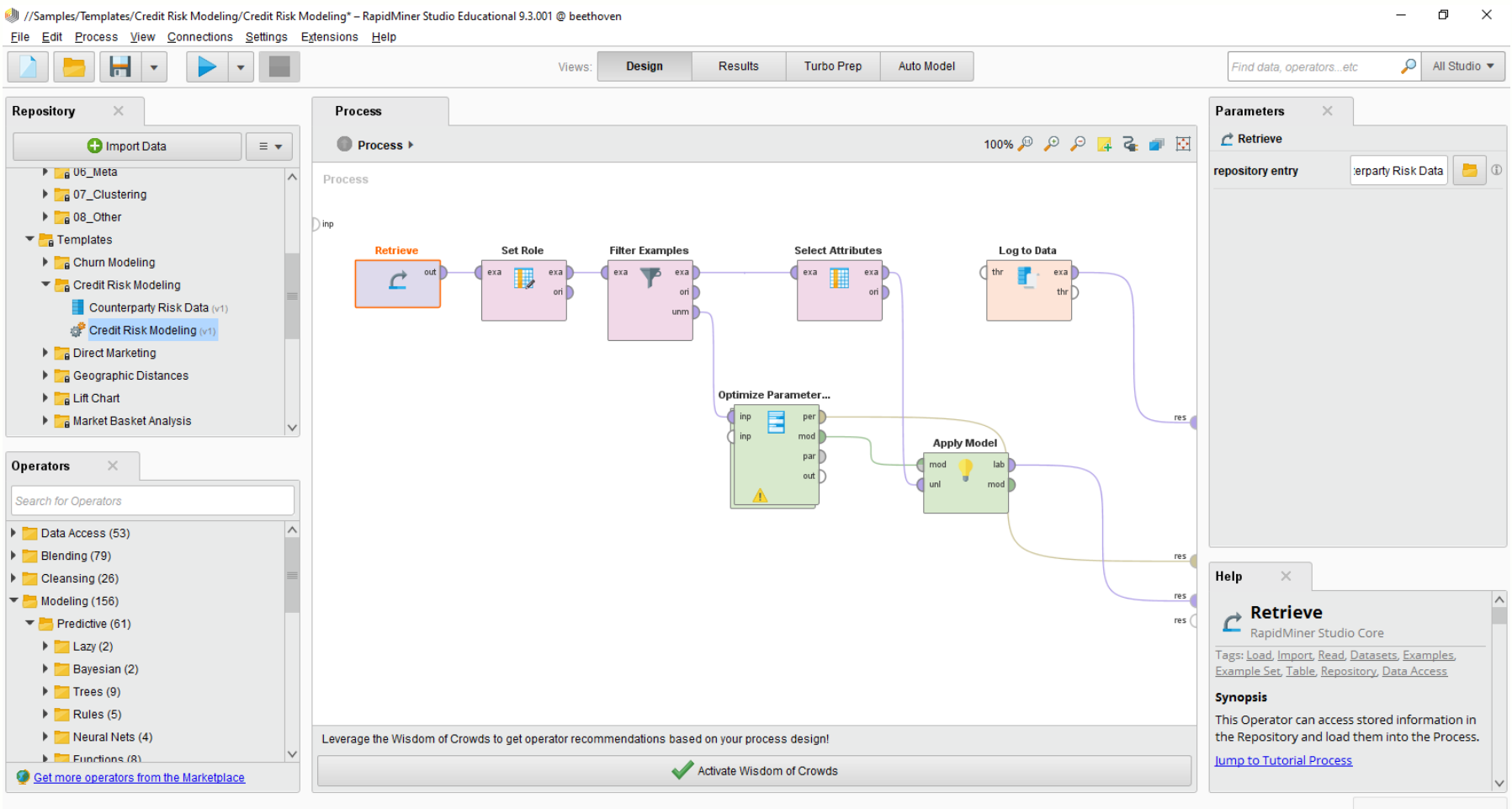
HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

- Umgebung für maschinelles Lernen und Data-Mining
- Programmiersprache: Java, damit weitgehend plattformunabhängig
- grafische Benutzeroberfläche
- mehr als 1500 Algorithmen des Maschinellen Lernens, über 30 interaktive Visualisierungen
- Austauschbarkeit von Datenvorverarbeitungsschritten, Lernverfahren und Evaluierungsverfahren
  - Durchführen von Vergleichen unterschiedlicher Verfahren
  - Kombination/Verschachteln von Verfahren
- RapidMiner Studio *Open Core* unter AGPL-Lizenz
- Free- (limitiert), Professional-, Enterprise-, Educational-Lizenzen

(*Visual Workflow for Predictive Analytics | RapidMiner© Studio* 2019; *RapidMiner Pricing* 2019)

RapidMiner

RapidMiner Studio



HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

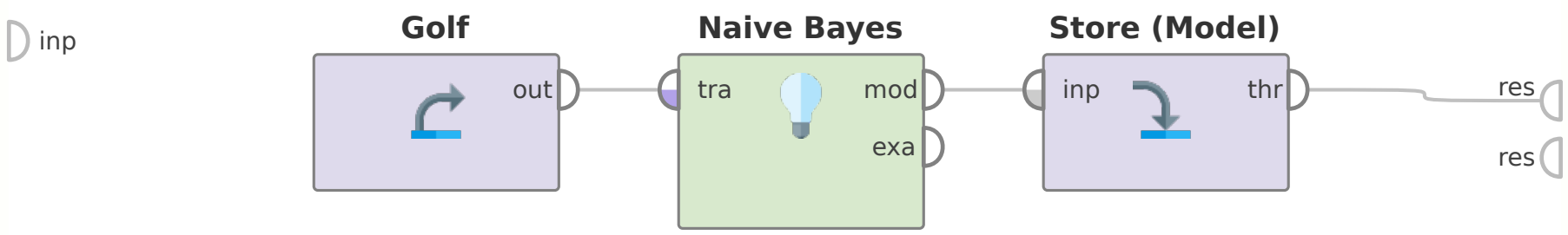
Grafische Benutzeroberfläche

- modularer Aufbau mit Knoten und einheitlichen Schnittstellen zwischen den Knoten die eine Erweiterung der Software ermöglichen
- Repository zur einheitlichen Speicherung von Datensets und Modellen
- Nachteil: geringere Performance
- Dashboard mit Operatoren, Repository, Prozess, Parametern, Daten-Editor und Hilfe-Seiten

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

Process

- ein Rapid-Miner-Prozess besteht aus einer Operatoren-Kette
- jeder Operator hat Parameter, die eingestellt werden können
- zwischen den Operatoren können sogenannte *IO Objects* (Input-Output-Objekte) weitergereicht werden
- die IO Objects können im internen *Repository* gespeichert werden



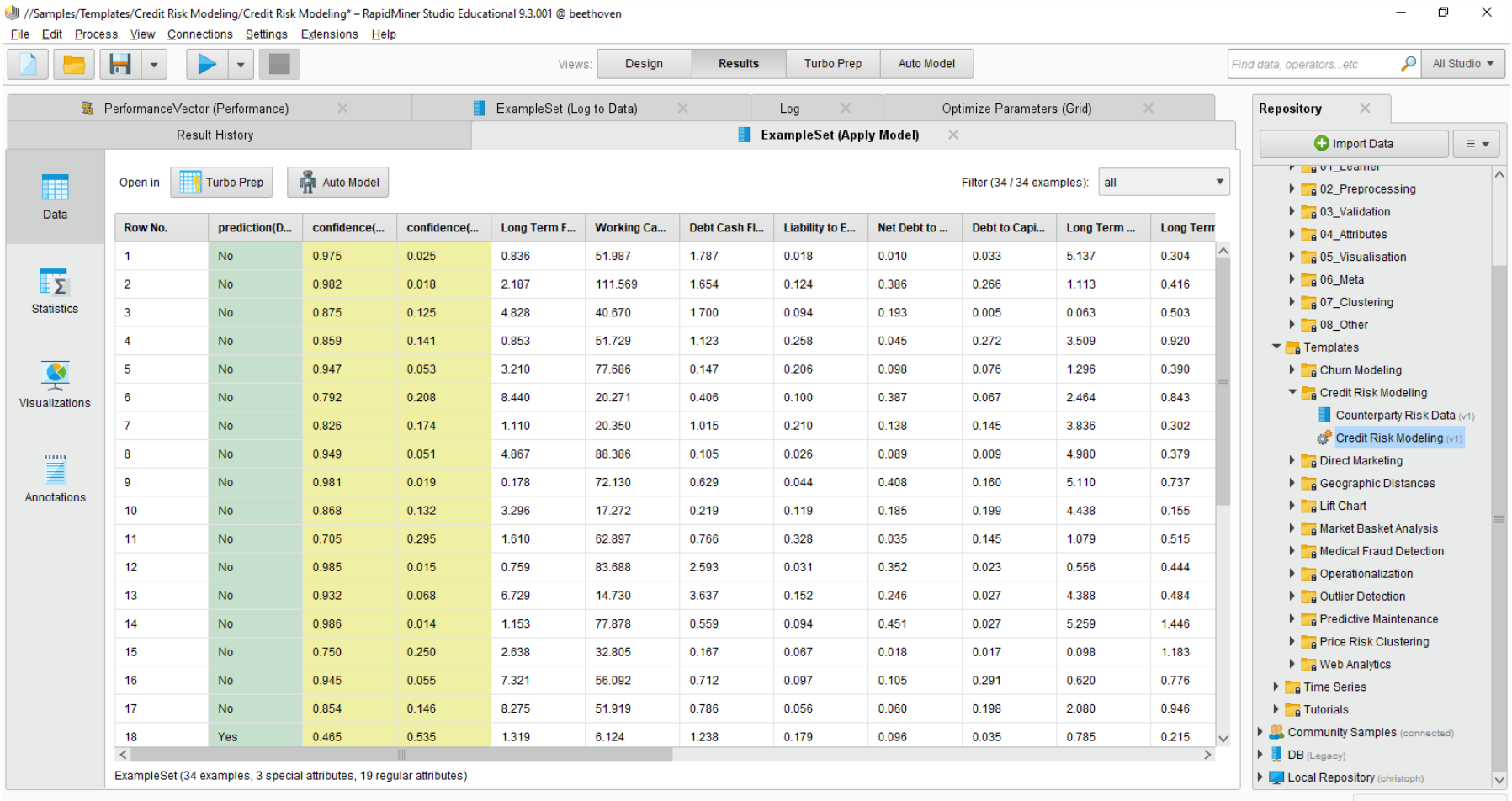
ExampleSet

- *ExampleSet* ist das Objekt, in dem Daten beschrieben und gespeichert werden
- ein ExampleSet ist eine Menge von Examples (Datensatz, math. Tupel)
- ein *Example* besteht aus einer Menge von Attributen
- jedes Example eines ExampleSets hat die gleichen Attribute
- ein ExampleSet ist im Prinzip eine Tabelle, die Zeilen sind die Examples und die Spalten die Attribute

RapidMiner

Funktionsweise

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES



RapidMiner

Attribute

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

1. Reguläre Attribute
   - aus diesen Daten wird das Modell erlernt
   - die Speziellen Attribute werden nicht für Lernoperationen verwendet
2. Spezielle Attribute
   - ID eindeutiger Identifier zum Identifizieren eines Examples; wird vor allem in der Datenvorverarbeitung zum Verbinden von Datensätzen (Join) benötigt
   - Label das Zielattribut welches vorhergesagt werden soll; für dieses Attribut wird das Modell trainiert
   - Prediction die aus dem Modell erstellte Vorhersage des Zielattributs
   - Cluster die Zugehörigkeit eines Examples zu einem Cluster
   - Weight das Gewicht eines Examples
   - Batch die Zugehörigkeit zu einem *Example Batch*

RapidMiner

Datentypen

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Value type	Name	Use
Nominal	nominal	Categorical non-numerical values, usually used for finite quantities of different characteristics
Numerical values	numeric	For numerical values in general
Integers	integer	Whole numbers, positive and negative Real numbers real Real numbers, positive and negative
Text	text	Random free text without structure
2-value nominal	binominal	Special case of nominal, where only two different values are permitted
multi-value nominal	polynominal	Special case of nominal, where more than two different values are permitted
Date Time	date_time	Date as well as time
Date	date	Only date
Time	time	Only time

(*RapidMiner Studio Manual* 2014)

Literatur I

Cleve, Jürgen und Uwe Lämmel (2014). *Data mining*. Studium. München: De Gruyter Oldenbourg. 306 S. ISBN: 978-3-486-72034-1 978-3-486-71391-6.

Ittner, Prof. Dr.-Ing. Andreas (2019). „Data Mining“. Vorlesung, Wintersemester 2019/20 (Hochschule Mittweida).

RapidMiner Pricing (2019). URL: https://rapidminer.com/pricing/ (besucht am 12. 09. 2019).

RapidMiner Studio Manual (2014). URL: https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf (besucht am 12. 09. 2019).

Visual Workflow for Predictive Analytics | RapidMiner© Studio (2019). URL: https://rapidminer.com/products/studio/ (besucht am 12. 09. 2019).