Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2019/20
IF17/MI17

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Datenvorverarbeitung

Sie haben bereits einige Faktoren kennengelernt, die Ihr Modell beeinflussen. Zur Erinnerung, dies waren insbesondere Ausreißer, Fehlende sowie falsch skalierte Werte. Die Behandlung dieser Einflussfaktoren lässt sich schwer automatisieren und wird deshalb in einem oder mehreren separaten Schritten nach den Vorgehensmodellen CRISP oder Fayyad ausgeführt (vgl. Cleve und Lämmel 2014, S. 4 – 10). Data Mining oder im Allgemeinen Maschinelles Lernen ist auf gute Daten angewiesen. Bevor belastbares Data Mining betrieben werden kann, ist die Qualität der Trainingsdaten zu untersuchen.

Aufgabe 1

Datenvorbereitung

a) Lesen Sie die Seiten 195 bis 201 in ebd.

b) Welche Arten der Datenvorbereitung gibt es?

c) Mit welchen Problemen sieht sich die Datensäuberung konfrontiert?

d) Wie können Sie auf fehlende Werte reagieren?

Aufgabe 2

Körperfettanteil mit RapidMiner

a) Informieren Sie sich über den Operator *Detect Outlier (Distances)*. Die Dokumentation finden Sie in *Detect Outlier (Distances) - RapidMiner Documentation* (2019).

b) Erstellen Sie einen neuen Prozess und importieren Sie die Datei Bodyfat.csv. Fügen Sie den Operator *Detect Outlier (Distances)* ein und lassen Sie sich das *ExampleSet* anzeigen. Führen Sie eine explorative Datenanalyse durch und achten Sie insbesondere auf Ausreißer. Erstellen Sie dazu verschiedene Scatter-Diagramme. Die *Information* (vgl. Cleve und Lämmel 2014, S. 38), ob es ich um einen Ausreißer handelt, kann durch die Farbe der Datenpunkte repräsentiert werden. Verwenden Sie einen sinnvollen Wert für die Anzahl der Ausreißer, vgl. Abbildung 1 und Abbildung 2. Probieren Sie auch verschiedene Werte für k aus.

c) Welchem Schritt im CRISP-Modell ist die Ausreißer-Erkennung zuzuordnen?

d) Entfernen Sie die Ausreißer sowie das Attribut *Dichte* und speichern Sie den Prozess sowie die verarbeiteten Daten im Repository.

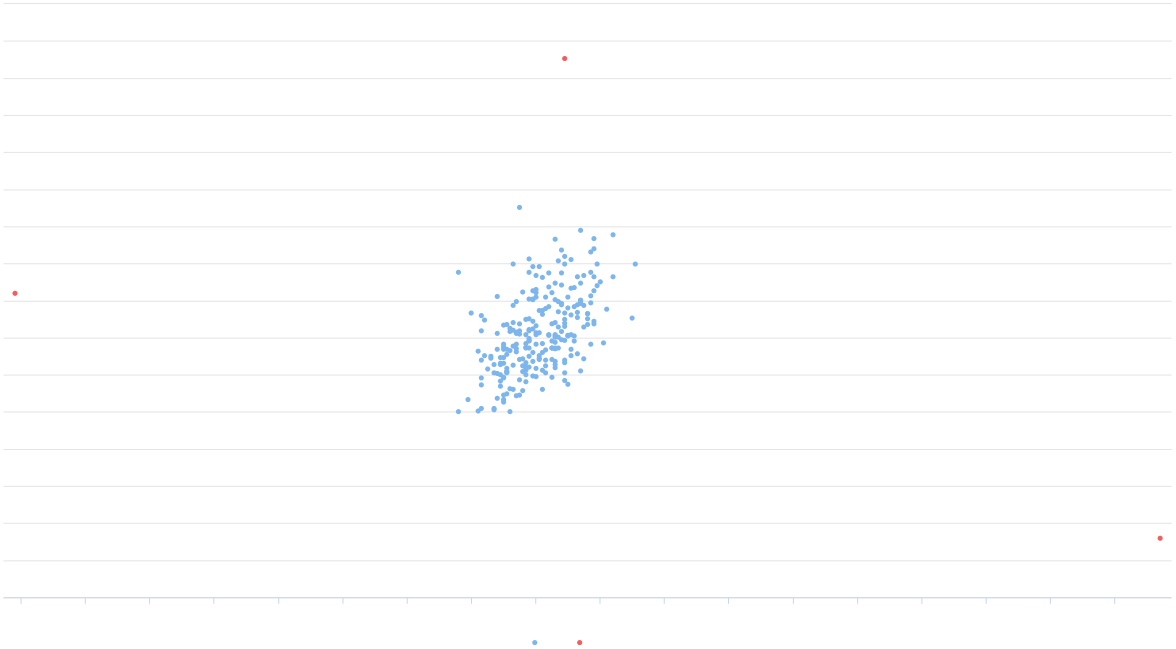


Abbildung 1: Ausreißer-Erkennung mit 3 Ausreißern; Attribute Körpergröße und Gewicht

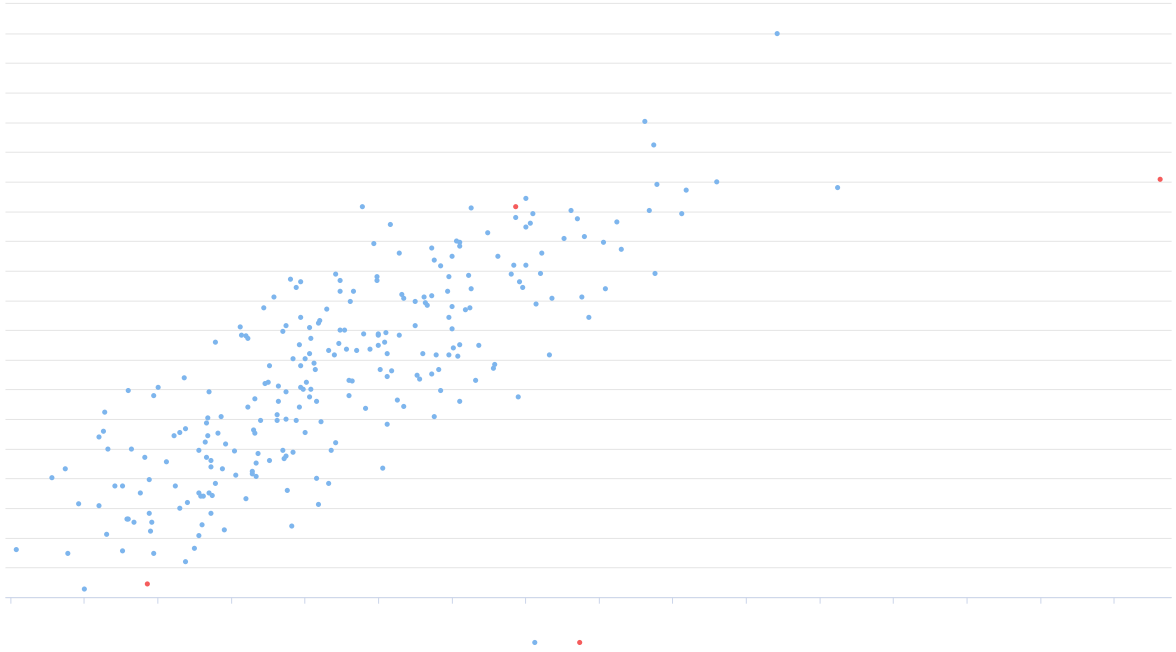


Abbildung 2: Ausreißer-Erkennung mit 3 Ausreißern; Attribute Körperfettanteil und Bauchumfang

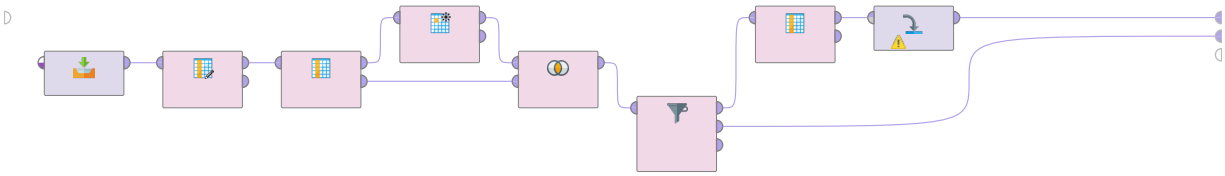


Abbildung 3: RapidMiner-Prozess zur Datenvorverarbeitung

e) Wenden Sie die Lineare Regression mit optimierter Attribute-Auswahl erneut an. Vergleichen Sie das Ergebnis mit dem aus den unbereinigten Daten.

Aufgabe 3

Sterberate mit RapidMiner

a) Sehen Sie sich die Daten in der Datei Death_rate.csv an. Was wird als Trennzeichen verwendet?

   Hinweis: Sie benötigen Reguläre Ausdrücke.

b) Erstellen Sie einen neuen Prozess zur Datenvorverarbeitung (DVV) und lesen Sie die Daten in RapidMiner ein. Führen Sie eine explorative Datenanalyse durch und achten Sie auf Ausreißer und fehlende Werte.

c) Lassen Sie die Ausreißer erkennen und entfernen Sie die betroffenen Datensätze. Entfernen Sie anschließend das Attribut *outlier* und speichern Sie den Prozess sowie die Daten im Repository (Abbildung 3).

Lösungen

1. b) Datenselektion und -integration, Datensäuberung, Datenreduktion, Datentransformation.

1. c) Fehlende Daten, Verrauschte Daten, Falsche Daten, Inkonsistente Daten.

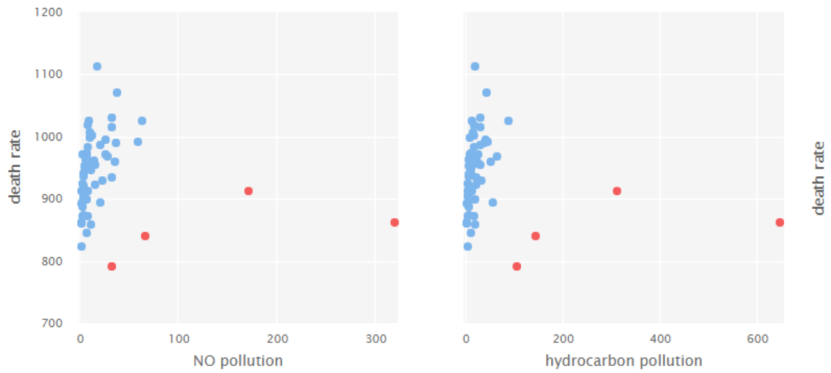
1. d) Attribut ignorieren, Fehlende Werte manuell einfügen, Globale Konstante, Durchschnittswert, Wahrscheinlichster Wert, Häufigster Wert, Relation zwischen Attributen, Datensatz als fehlerhaft kennzeichnen.

2. b) Detect␣Outlier.rmp

2. c) Schritt 3: Datenvorbereitung

3. a) $1 - n$ Leerzeichen; RegEx: \s+

3. b) Es gibt 4 auffällige Ausreißer bei den Attributen „hydrocarbon pollution“ und „NO pollution“. Fehlende Werte sind nicht vorhanden.



Literatur

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.

Detect Outlier (Distances) - RapidMiner Documentation (2019). URL: https://docs.rapidminer.com/latest/studio/operators/cleansing/outliers/detect_outlier_distances.html (besucht am 11. 10. 2019).