Logistische Regressionsanalyse

Referenten: Sodaba Hayat

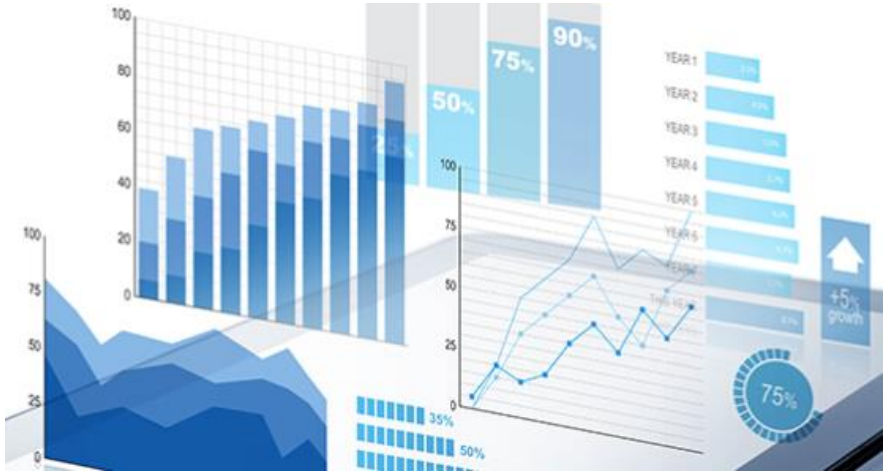
& Sahand Armin

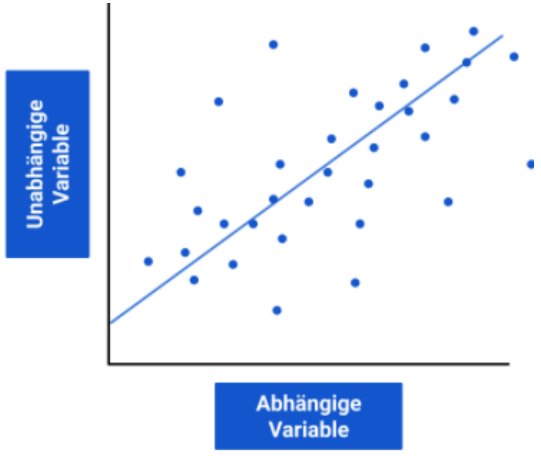
Inhaltsverzeichnis

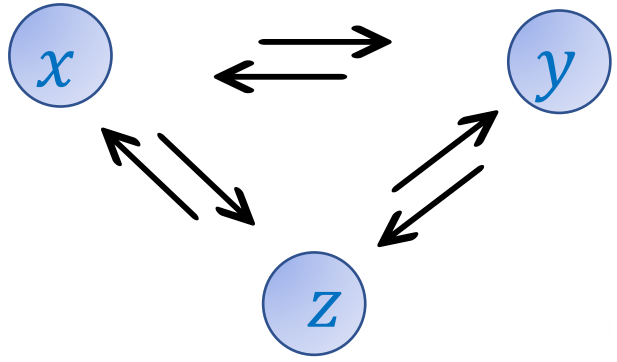
Allgemein zu Regressionsanalyse

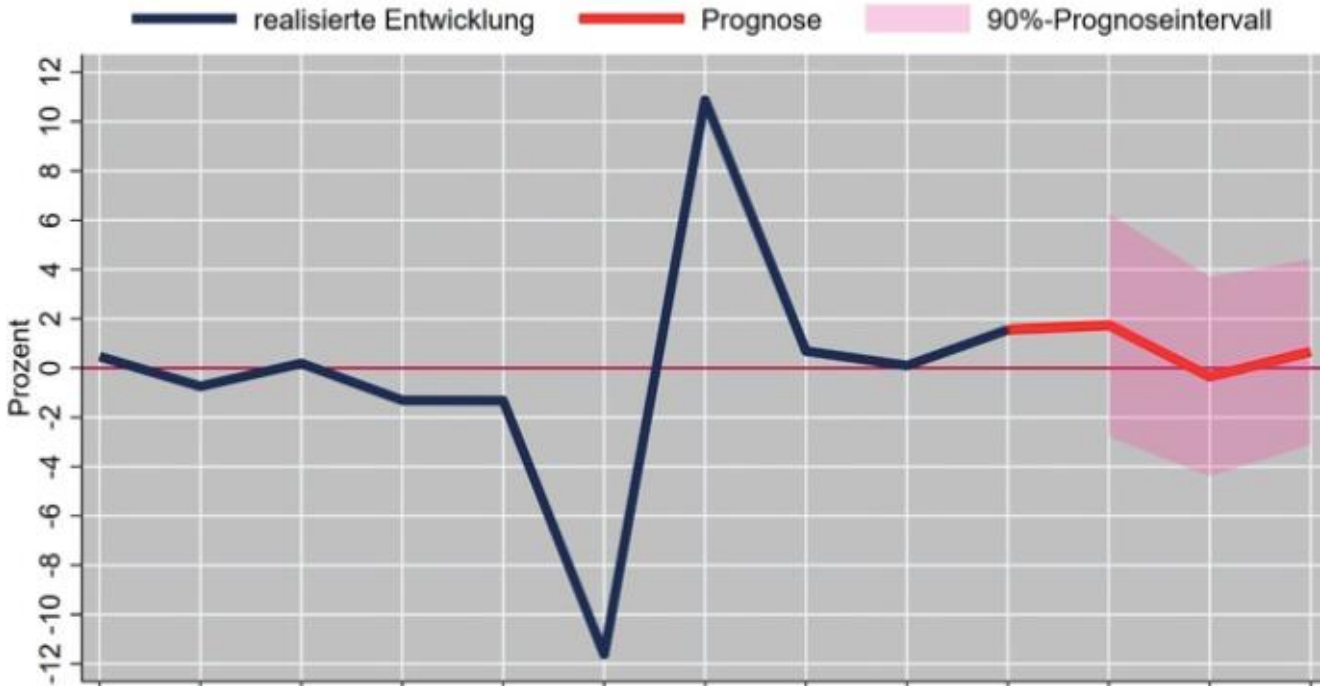
- Statistisches Analyseverfahren

➢ Zusammenhang zw. 2 oder mehr unabhängigen oder abhängigen Variablen









Erstellung von Vorhersagefunktion & Vorhersage von neuen Werten

Arten von Regressionsanalyse

1. Einfache lineare Regression
2. Multiple lineare Regression
3. Logistische Regression
4. Multivariate Regression

Lösen von Regressionsprobleme

Lösen von Klassifikationsprobleme

Logistische Regressionsanalyse Einführung

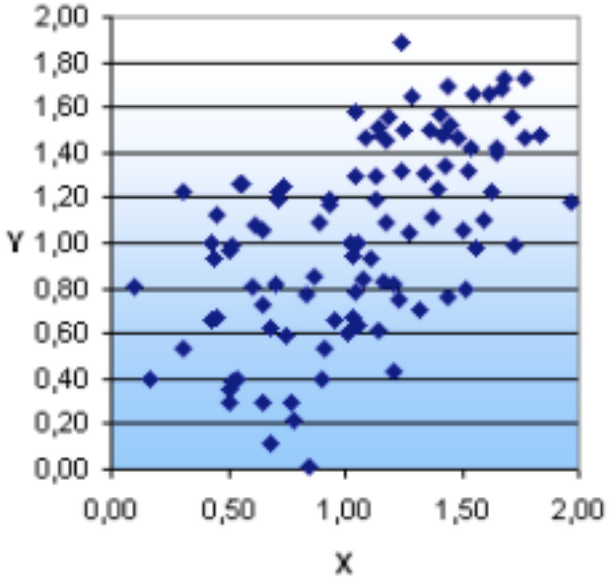
1…n unabhängig — ZUSAMMENHANG zw. VARIABLEN — 1 abhängige Zielvariable

Prädiktor

x

- erklärend

→ Alter, Geschlecht, Rauchen



Kriterium

y

- kategorisch,
- nominalskaliert

binär

2 Ausprägungen

JA
Wahr
1

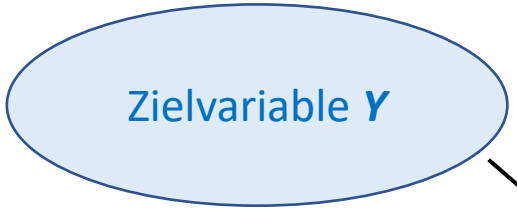
NEIN
Falsch
0

multinomial

Mehr als 2 Ausprägung

Motivation

- Einflüsse der diskreten(unabhängigen) Variablen mit linearer Regression unmöglich



Lineare Regression

metrisch oder intervall skaliert

$(y = 1, 2, 3, 4 \ldots \ldots)$

Logistische Regression

kategorische Variable

nominal-, ordinalskaliert

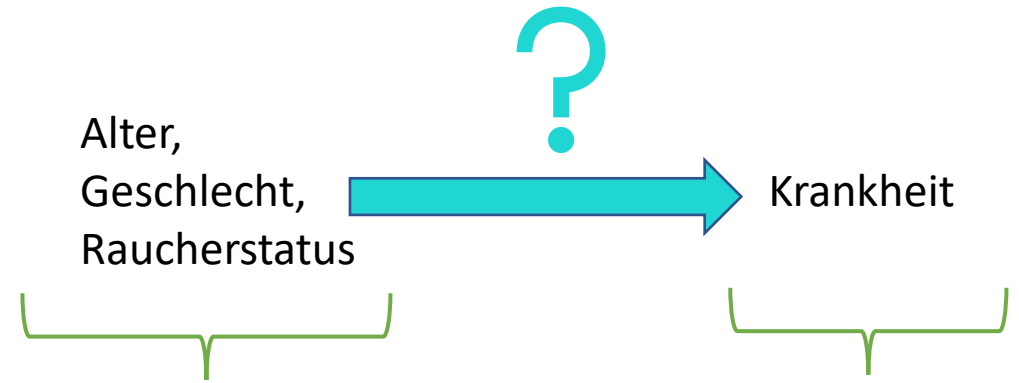
$(y = 0;\ y = 1)$

Binäre Logistische Regression

Ausgabe der Berechnung:

- keine konkrete Werte
- ***Y-Werte*** nur ***1*** & ***0***
- Berechnung der Wahrscheinlichkeit ***P*** für 2 Ausprägungen der ***Y***

$$0 < P < 1$$
$$P(JA) = 80\% = 0{,}8$$
$$P(NEIN) = 20\% = 0{,}2$$



P(Y=1/ JA)

P(Y=0/ NEIN)

krank

Nicht krank

Mathematische Funktion & Kurve der binären Logistischen Regression

$$p(y=1) = \frac{1}{1+e^{-z}}$$

$$z = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k + \varepsilon$$

$$p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 * x_1+\beta_2 * x_2+\cdots+\beta_k * x_k+\varepsilon)}}$$

z : Logit: lineares Regressionsmodell der abhängig. Variablen

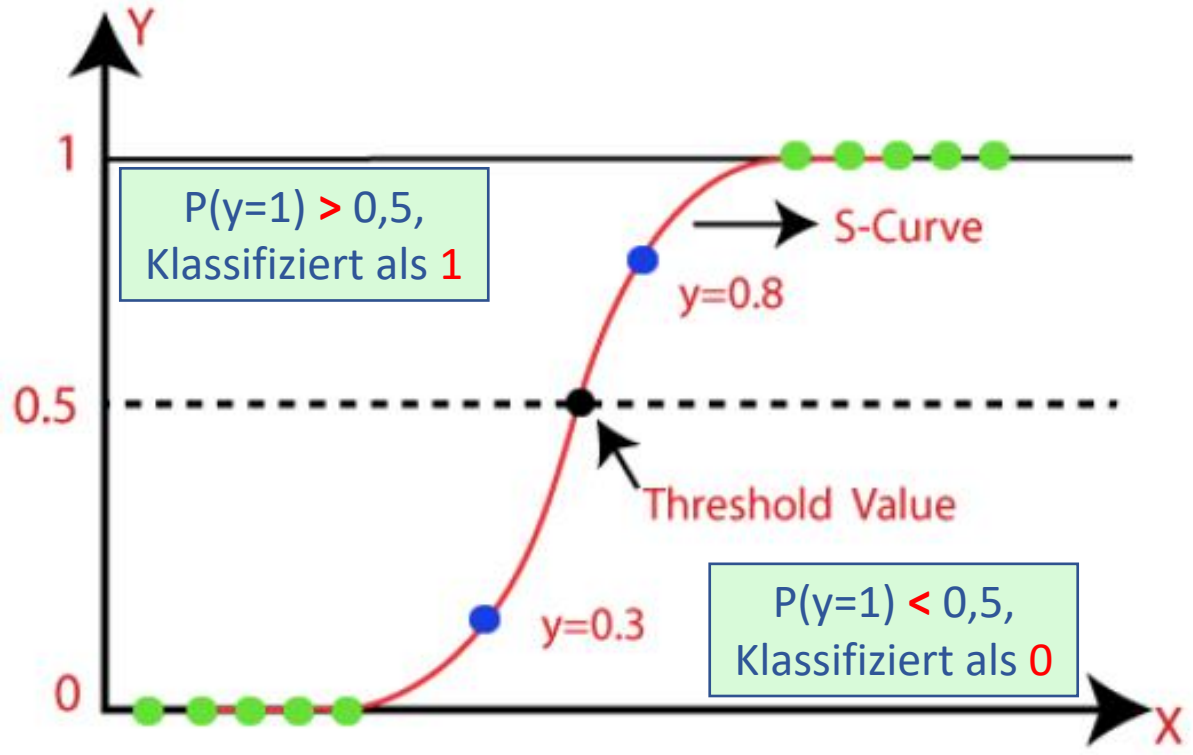
x : Unabhängig. Var.

β_k: Regressionskoeffizient

ε : Fehlerwert

Sigmoide Funktion

2 Maximalwerte: **0** oder **1**

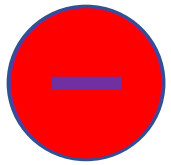


Maximum Likelihood Schätzung

Schätzung der Regressionskoeffizienten β

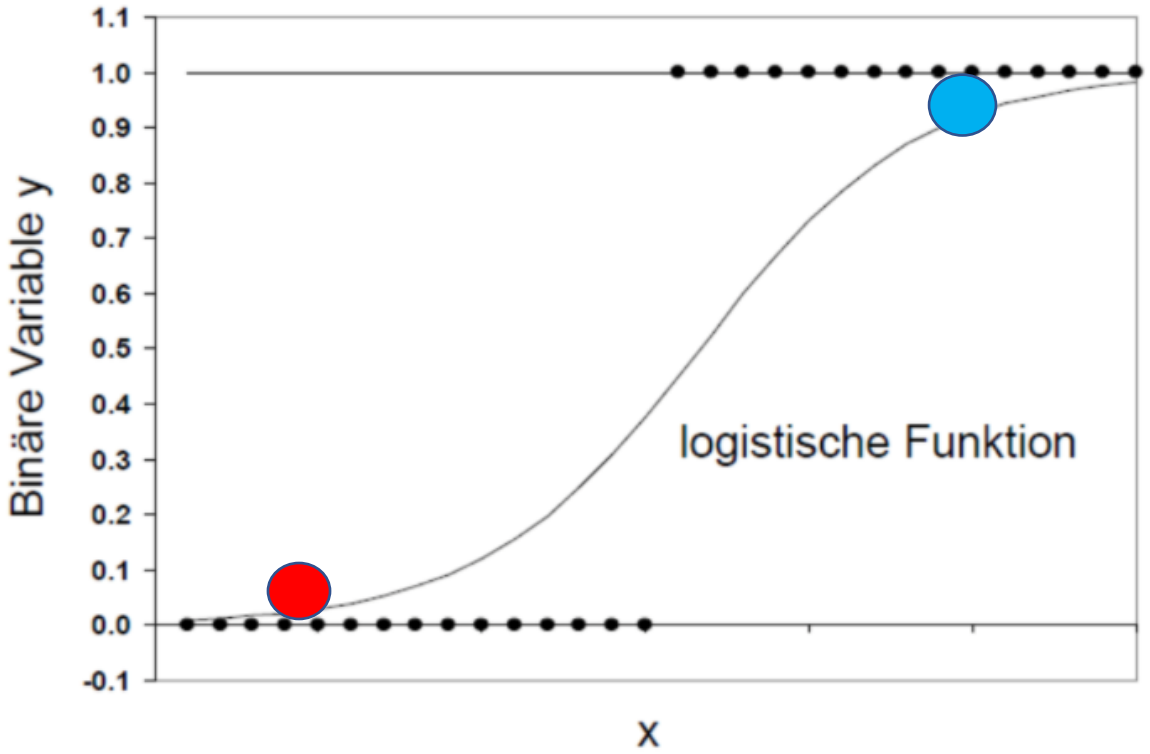
β:

oder



P(y=1) nahe 1, Je höher Prädiktor

P(y=1) nahe 0, Je höher Prädiktor



Anwendungsvoraussetzungen

y Abhängige Variable

Binär kodiert
0,1

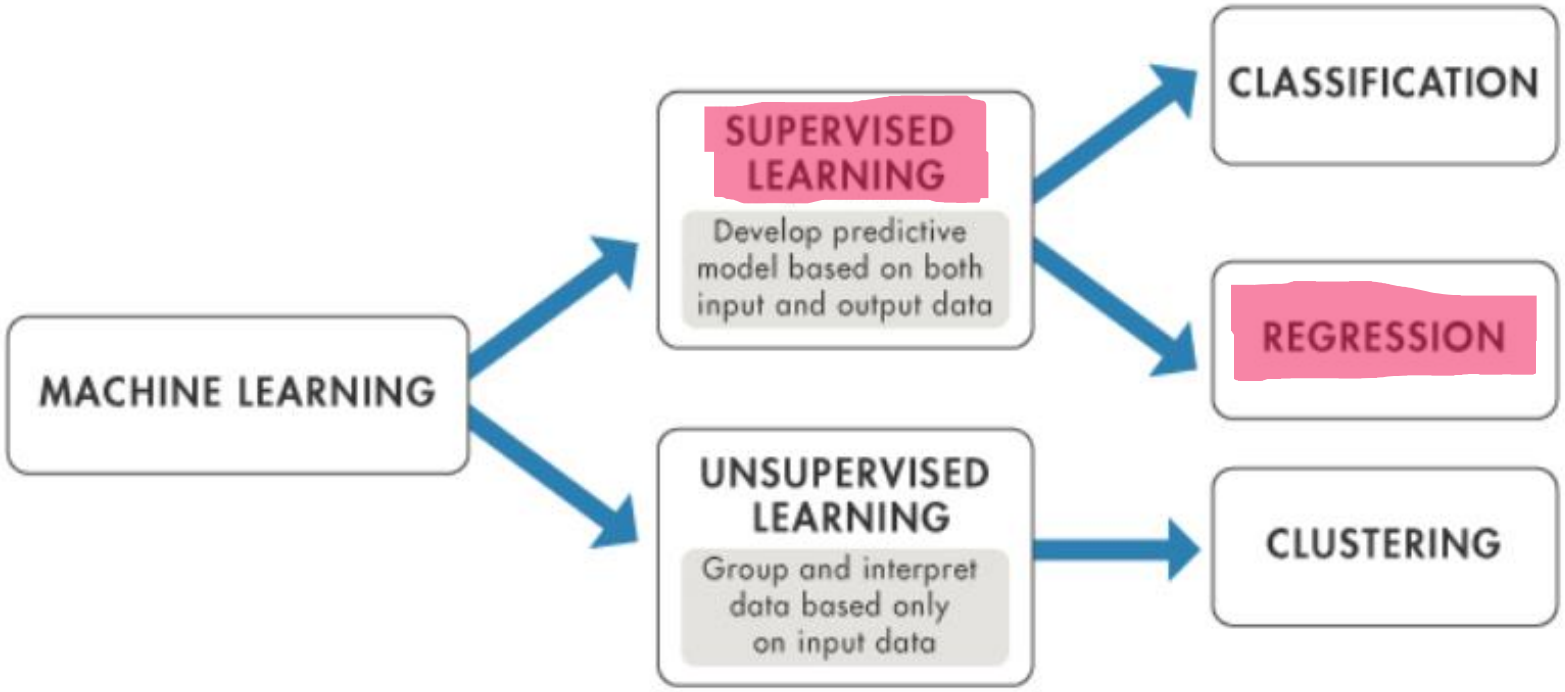
---

x Unabhängige Variable

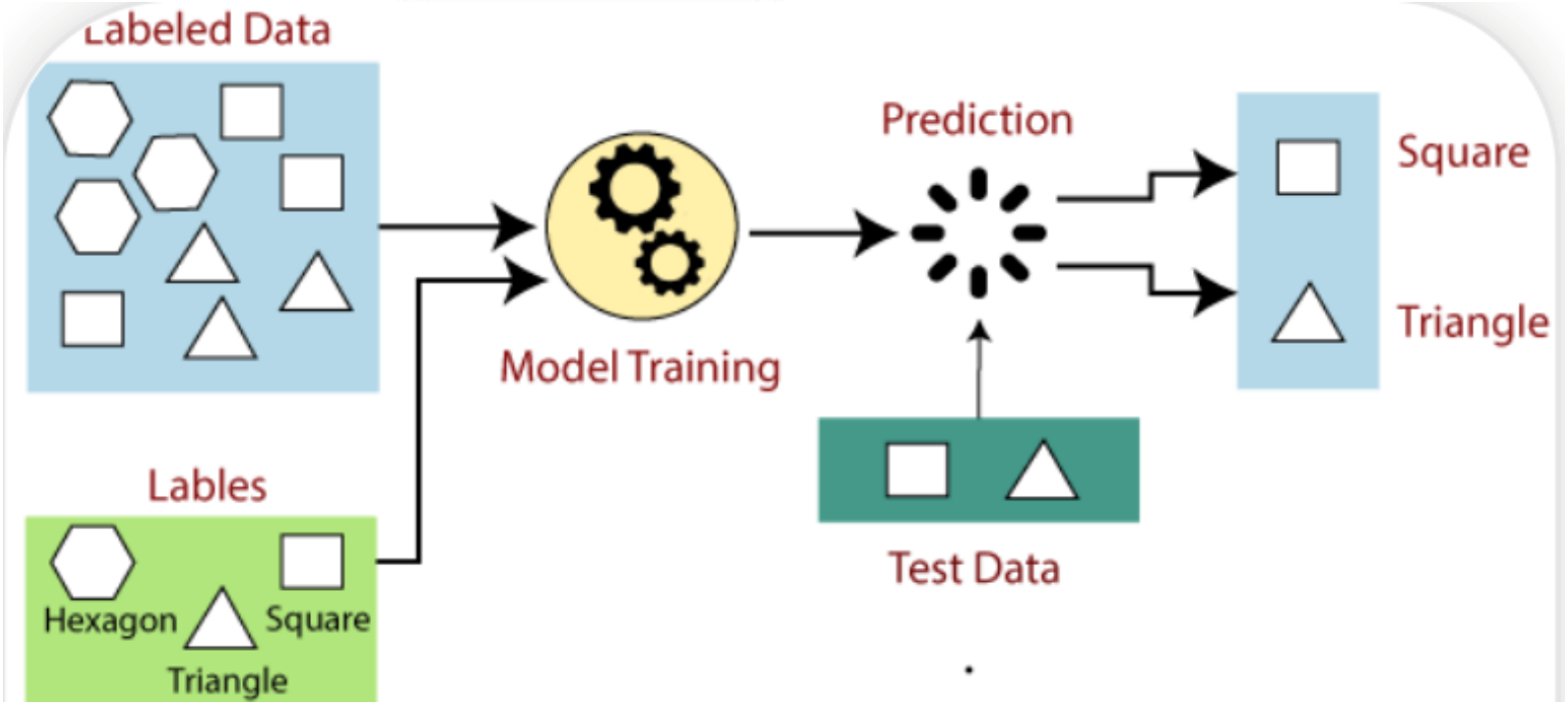
Kein Zusammenhang untereinander

Metrisch oder als Dummy Variablen kodiert (kategorisch)

Stichprobe
$n \geq 25$
(kategorisch)

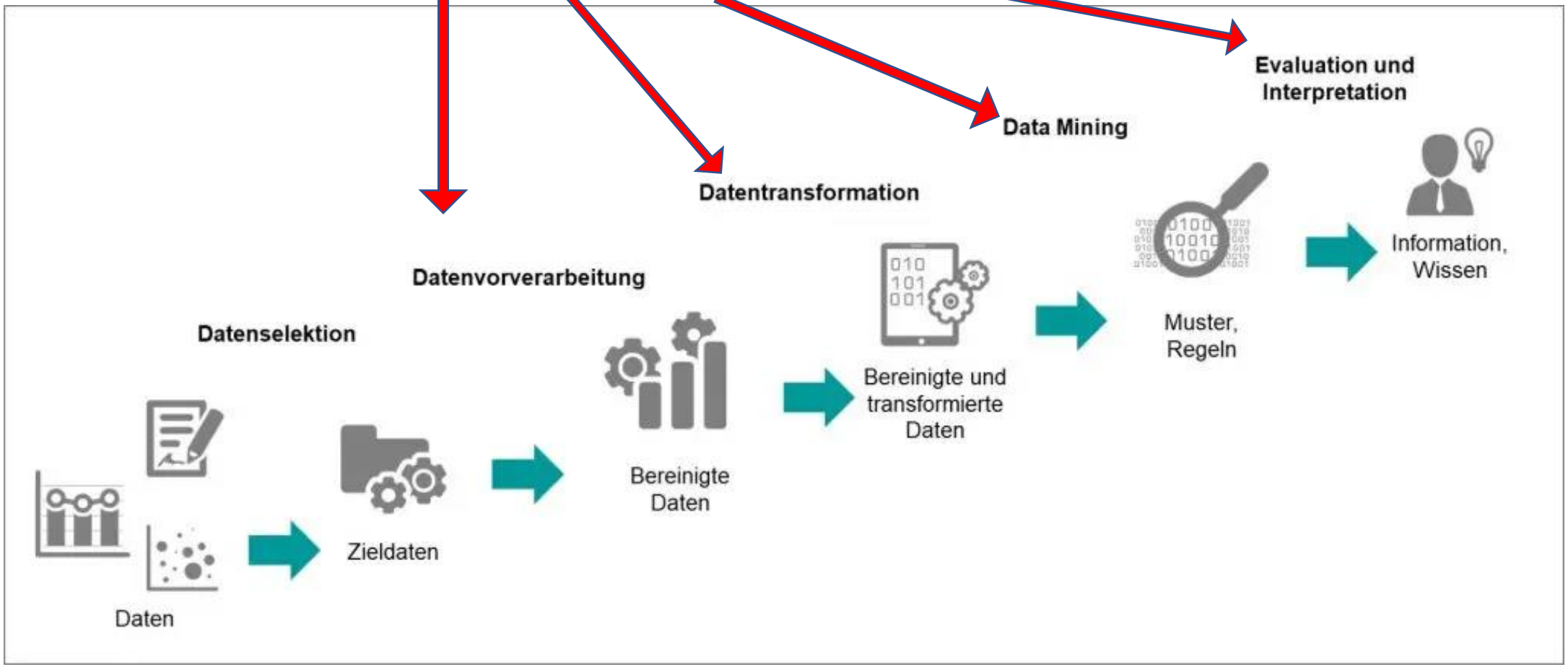


Supervised Learning



Einordnung in Phasen des KDD

Logistische Regression

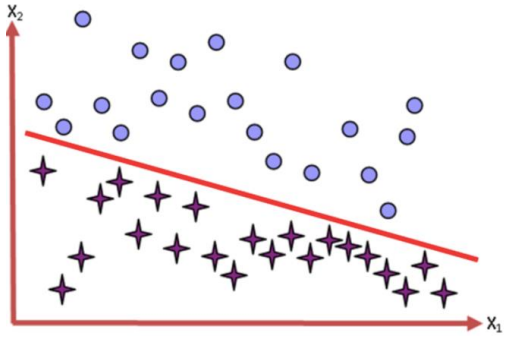


Vorteile

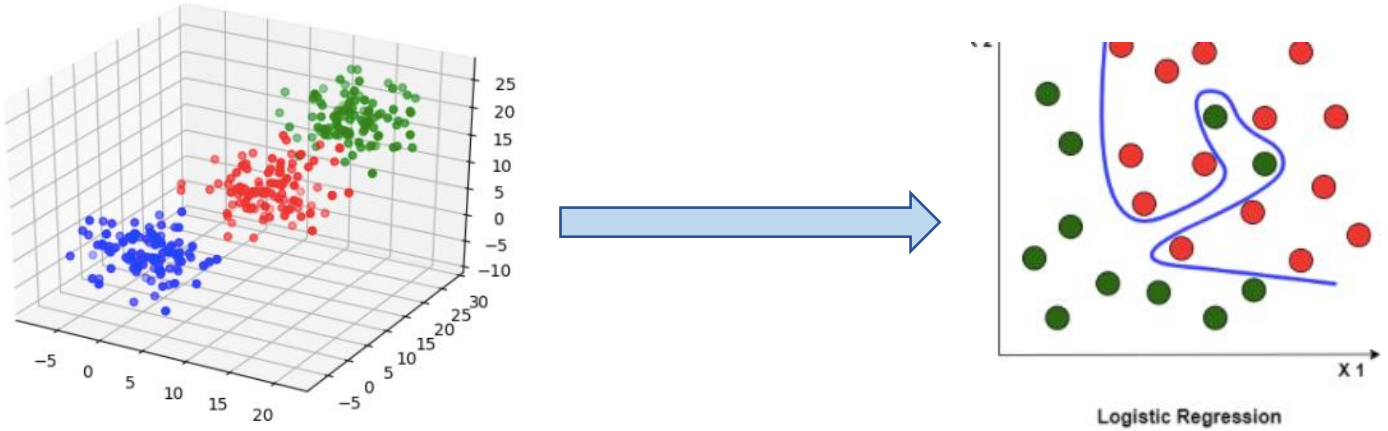
- Klassifikationsmodell & **gibt Wahrscheinlichkeit (P) an**
- Einfacher zu implementieren, interpretieren & effizienter zu trainieren
- Sehr effizient, wenn der Datensatz linear trennbare Merkmale aufweist
- weniger anfällig für Überanpassungen in einem niedrigdimensionalen Datensatz mit ausreichend Trainingsbeispielen

Nachteile

- Kann nicht mehr trainiert werden,
wenn es ein Merkmal gibt, das die beiden Klassen perfekt trennt



- Überanpassungen (Outfitting) bei hochdimensionalen Datensätzen.

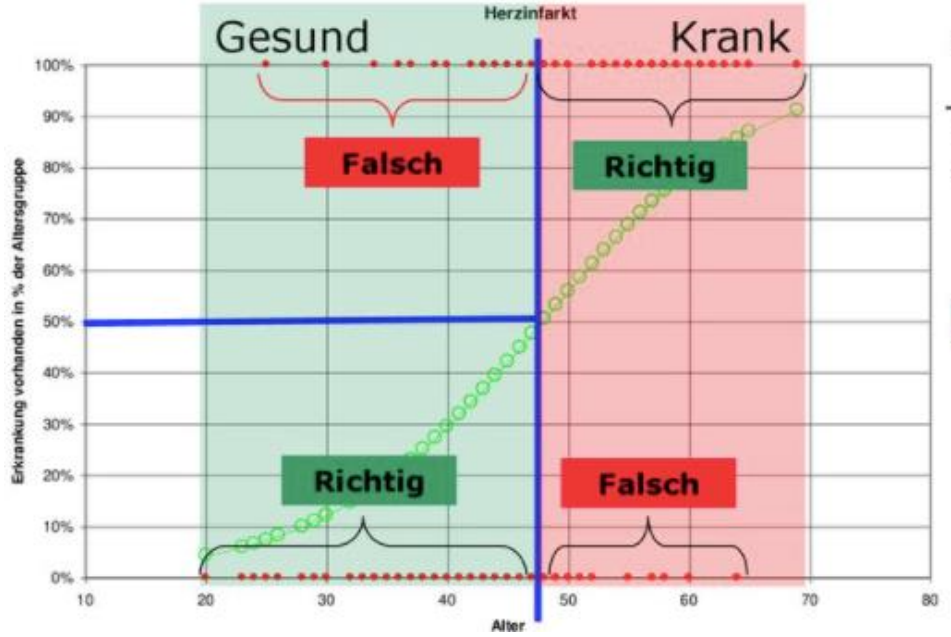


- Nicht für komplexe Beziehungen geeignet → Neuronale Netze

Anwendungsgebiete

Marketing, Human Ressources, Finanzen, Medizin, Wissenschaft Forschungspraxis…….etc.

- **Wissenschaft :** Erdbeben
- **Medizin :** Prognose über zukünftiger Verlauf einer Krankheit



Klassifizierungstabelle

		Prognose Gesund	Prognose Krank	% Richtig
Beobachtet	Gesund	45	12	79%
	Krank	14	29	67%
Gesamt		59	41	74%

Insgesamt werden 74% der Probanden richtig klassifiziert, die Gesunden etwas besser als die Kranken.

Trefferrate = 74%

Implementierung in Python

1. Datenvorverarbeitung
2. Anpassen der Logistischen Regression an das Trainingset
3. Vorhersage des Testergebnisses
4. Testgenauigkeit des Ergebnisses(Erstellung von Confusion Matrix)
5. Visualisierung des Testergebnisses

Schulung:

- CSV- Datei
- Notebook
- Cheat Sheet

Fragen?

Der Link für Quiz wird Ihnen im Chat zur Verfügung gestellt

Vielen Dank für Ihre Aufmerksamkeit

Quellenverzeichnisse

http://archiv.ub.uni-heidelberg.de/volltextserver/4073/1/Diplomarbeit_Christian_Gottermeier.pdf

https://www.iat.eu/aktuell/veroeff/2003/erling07.pdf

https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/zusammenhaenge/lreg.html

https://ichi.pro/de/logistische-regression-131466354476160

https://www.sowi.uni-stuttgart.de/dokumente/forschung/siss/2010.SISS.3.pdf

https://www.amazon.de/Entwicklung-Validierung-Prognosemodellen-logistischen-Regression/dp/3832235272

https://www.javatpoint.com/logistic-regression-in-machine-learning

Bildquellenverzeichnis

https://de.m.wikipedia.org/wiki/Datei:Part_korrelation.PNG

https://morethandigital.info/grundlagen-des-data-mining-ein-prozess-ueberblick/

https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/zusammenhaenge/lreg.html

https://www.fapgf.top/ProductDetail.aspx?iid=59184092&pr=42.88

https://www.dpma.de/dpma/veroeffentlichungen/statistiken/index.html

https://www.presseportal.de/pm/118695/5033058

https://www.javatpoint.com/logistic-regression-in-machine-learning