



KLASSIFIZIERUNG DES SUIZIDRISIKOS VON PERSONENGRUPPEN IN LÄNDERN EUROPAS

- TEAM 1 -

OUR VISION & MISSION

weltweit eine der
10 häufigsten
Todesursachen

alle
40 Sekunden
nimmt sich ein Mensch
das Leben



**Jede
Altersgruppe**
ist betroffen

weltweit jährlich
> 800.000
Suizide



SOCIAL4SUICIDE E.V.



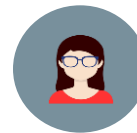
Sebastian
Wölk



Andreas
Herb



David
Schall



Lara
Müller

Das Team

ZIELGERICHTETE REDUZIERUNG DER
SUIZIDRATE INNERHALB EUROPAS
DURCH KLASSIFIZIERUNG DES
SUIZIDRISIKOS LÄNDERBEZOGENER
PERSONENGRUPPEN

OUR TASK



Fragestellungen

- Welche länder- und personenspezifischen Kriterien haben Einfluss auf die Suizidrate?
- Welche Personengruppe bekommt Unterstützung? (suizidkritisch/ unkritisch)
- Wie sieht diese Unterstützung aus? (z.B. finanziell, Altersgruppe, Geschlechter)



Zielstellung

- Aussprechen von Handlungsempfehlungen für Politik & Länder
- Entscheidung über Budgetverteilung (z.B. Beratungseinrichtungen)



Prämissen

- Betrachtung von Europa
- Betrachtung auf Länderebene & Bevölkerungsgruppen anstelle von einzelnen Individuen



Erfolgskriterien

Mindestanforderung an Vorhersagegenauigkeit:

Accuracy \geq 80%

→ Mindestens 80% aller Risikogruppen müssen richtig klassifiziert werden (TP & TN), um den Algorithmus für den Use Case einsetzen zu können

Im Entscheidungsfall wird der Algorithmus mit den **höchsten Recall** Werten bevorzugt.



DOMAIN EXPLORATION

- ANDREAS HERB, DAVID SCHALL -

DOMAIN EXPLORATION: SUIZID & DATA SCIENCE

Individuelle Suizidprävention

- **Ziel:** Identifikation von Risikogruppen /-menschen mit Data Science
- **Daten:** individuell auf einzelne Personen, vergangenheitsbezogen oder realtime
- **Methoden:** text mining & natural language processing

Gesellschaftliche Suizidprävention

- **Ziel:** Explorative Analyse & Relation von gesundheitlichen Faktoren (z.B. Depression, Alkohol, Drogen) auf Suizidraten
- **Daten:** Landesebene
- **Methoden:** explorative Analysen, z.T. Regressionsmodelle

Methodenfokus:

Inhaltsfokus:

BISHER

→ Regression & Text Mining

→ Individuum & Gesundheit

NEU

→ Klassifikation

→ Wirtschaft & Sozioökonomie





DATA EXPLORATION

- SEBASTIAN WÖLK, LARA MÜLLER -

ALLGEMEINES - DATENSATZSTRUKTUR

Suicide Rates 1985-2016

Gejointe Datensätze

- Allgemeine Daten
(z.B. Country-Codes, Kontinente)
- Ökonomische Daten
(z.B. Inflation, Gini, Employment etc.)
- Gesundheitsdaten
(z.B. Lebenserwartung, Investitionen in Gesundheitswesen etc.)
- Gesellschaftliche Daten
(z.B. HDI, Demokratieindex, Mordfälle etc.)

5088
Beobachtungen

28 Variablen
19 numerisch | 9 nominal

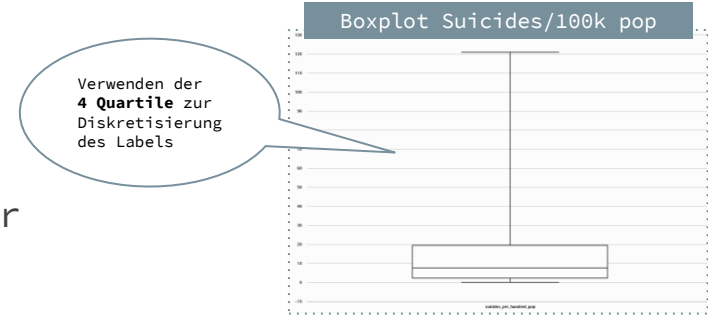
Kontinent:
Europa

Jahre: 2005-
2015



PREDIKTORVARIABLE / LABEL

- Zielvariable: Suizide pro 100.000 Einwohner
 - kategoriale Ausprägungen
 - **no risk** (≤ 2.09 Suizide pro 100.000)
 - **low risk** (> 2.09 und ≤ 7.66 Suizide pro 100.000)
 - **medium risk** (> 7.66 und ≤ 19.37 Suizide pro 100.000)
 - **high risk** (> 19.37 Suizide pro 100.000)



Index	Nominal value	Absolute count	Fraction
1	no_risk	1276	0.251
2	medium_risk	1272	0.250
3	high_risk	1271	0.250
4	low_risk	1269	0.249

NOMINAL: GESCHLECHT, ALTERSGRUPPEN, GENERATIONEN

Geschlecht

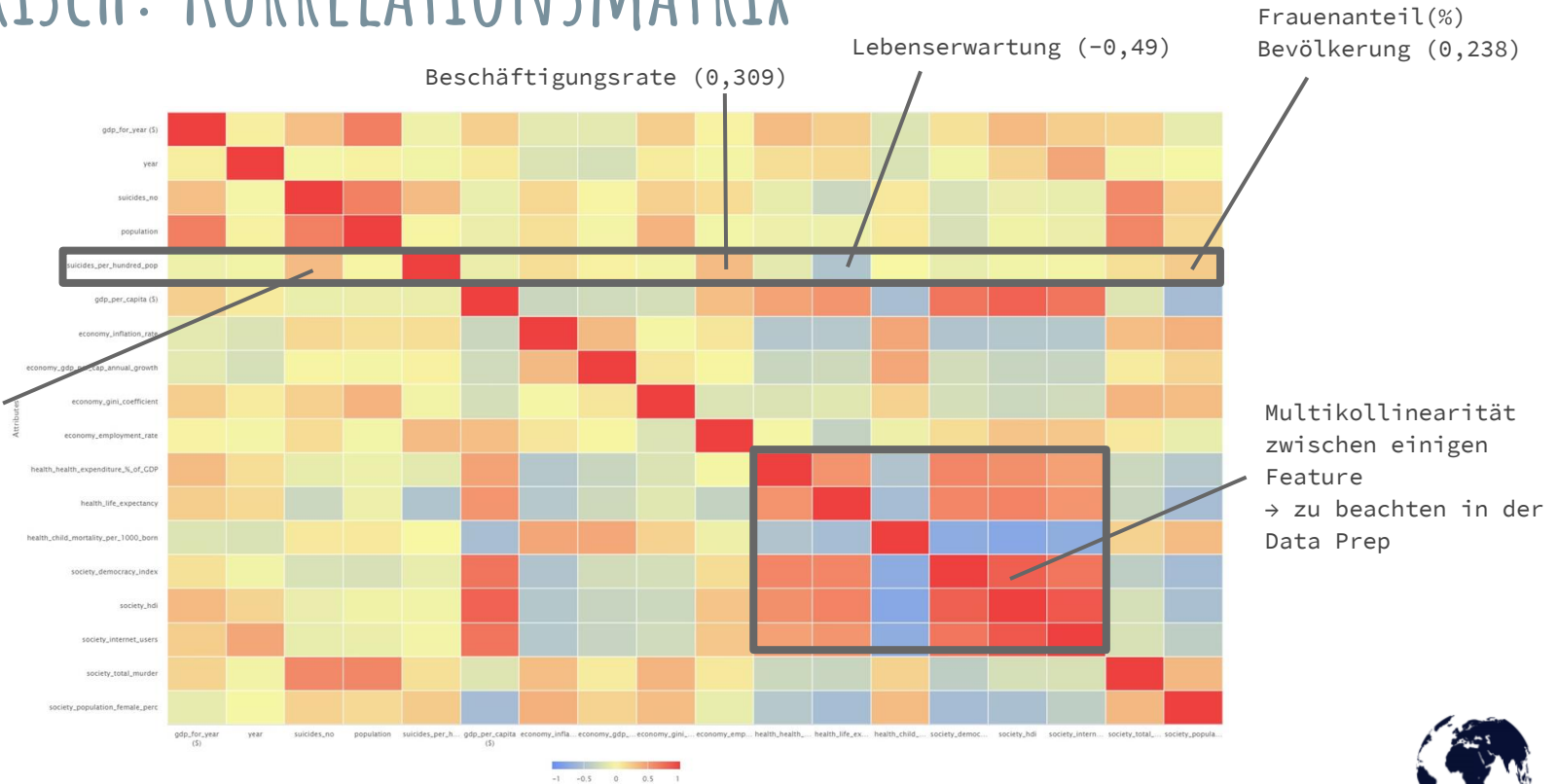
Index	Nominal value	Absolute count	Fraction
1	female	2544	0.500
2	male	2544	0.500

Altersgruppe

Index	Nominal value	Absolute count	Fraction
1	15-24 years	848	0.167
2	25-34 years	848	0.167
3	35-54 years	848	0.167
4	5-14 years	848	0.167
5	55-74 years	848	0.167
6	75+ years	848	0.167

Index	Nominal value	Absolute count	Fraction
1	Millenials	1374	0.270
2	Silent	1320	0.259
3	Generation X	928	0.182
4	Boomers	768	0.151
5	Generation Z	698	0.137

NUMERISCH: KORRELATIONSMATRIX

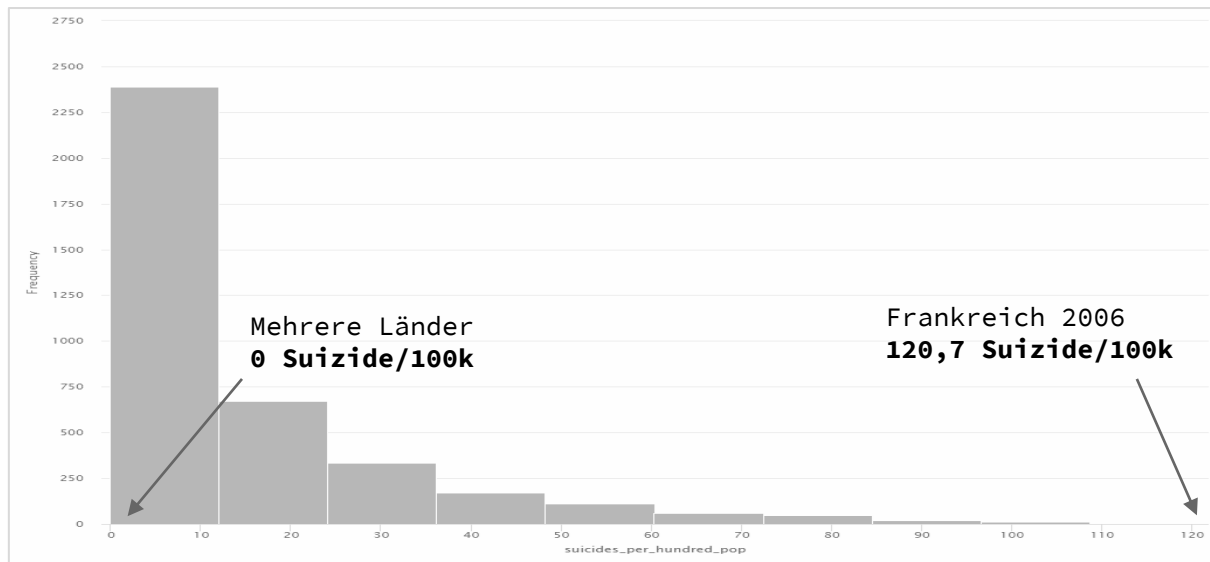


FEHLENDE WERTE / UNVOLLSTÄNDIGKEITEN

- Zwei Attribute mit fehlenden Werten
 - “Health expenditure in % of GDP” - 132 fehlende Werte
 - “Democracy index” - 552 fehlende Werte

✓ health_health_expenditure_%_...	Real	132	Min 2.690	Max 11.900	Average 7.891
✓ society_democracy_index	Real	552	Min 0.295	Max 0.993	Average 0.752

DESKRIPTIVE STATISTIK - SUIZIDE PRO 100K POPULATION



INFO

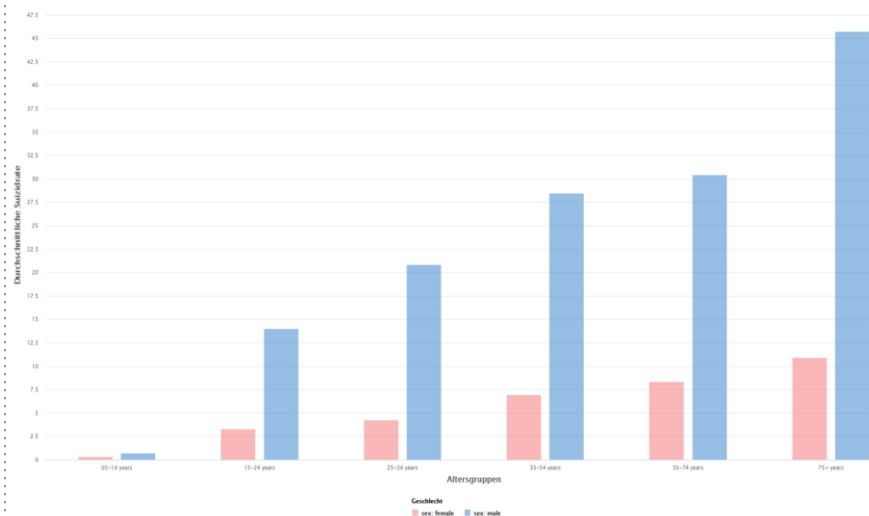
- Stark rechtsschiefe Verteilung der Suizidrate mit einer Range von 0-121 Suizide pro 100k Einwohner
- Viele Länder mit geringer, wenige mit einer sehr hohen Suizidrate

Avg	Min	Max	Median	Modus	Std. Dev.
14.396	0	120.750	7.660	0	18.228

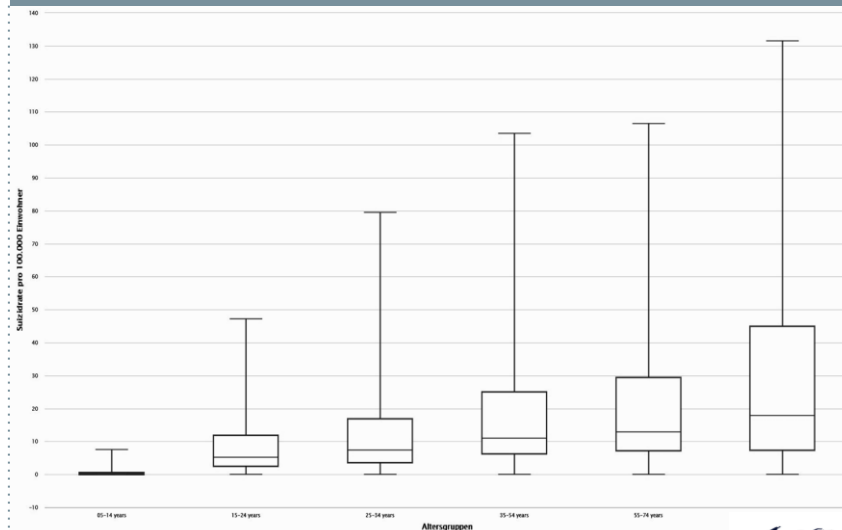
ALTERSGRUPPE - MEHR SUIZIDE IM HOHEN ALTER

Generation X
Silent
Boomers
Millennials

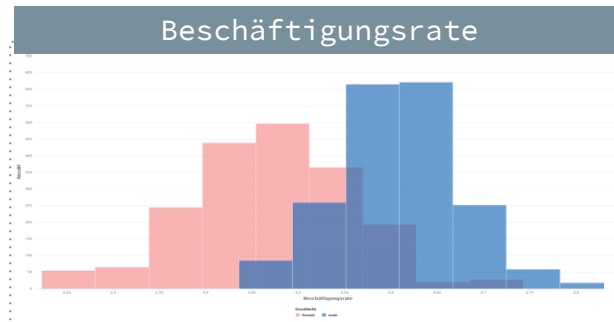
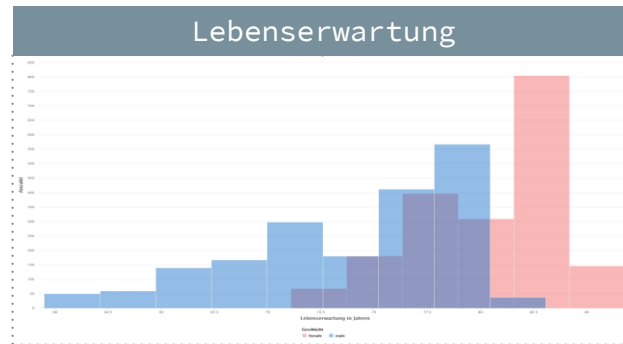
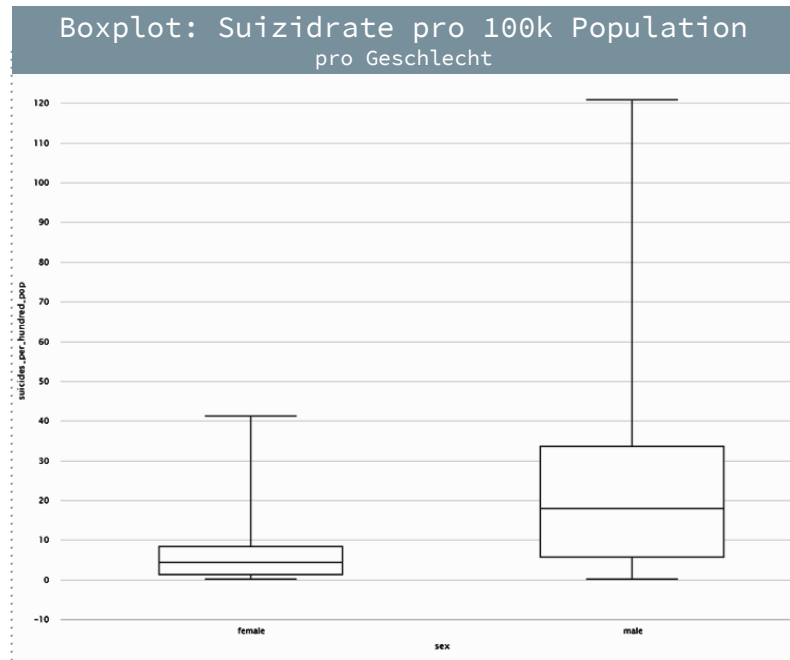
Suizidrate pro 100k Population
pro Altersgruppe & Geschlecht



Boxplot: Suizidrate pro 100k Population

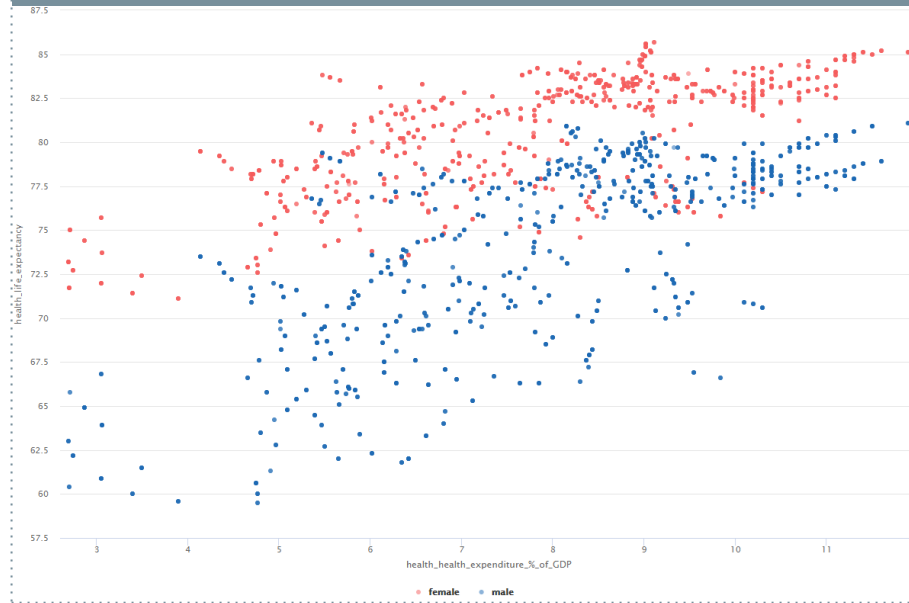


GESCHLECHT: MEHR SUIZIDFÄLLE BEI MÄNNERN



LEBENSERWARTUNG

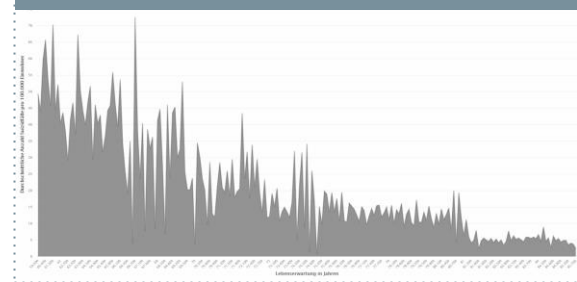
Lebenserwartung in Abh. zu Gesundheitsausgaben
pro Geschlecht



INFO

- Mit steigenden Ausgaben im Gesundheitssystem steigt auch die Lebenserwartung der Menschen
- Höhere Lebenserwartung führt zu einer geringeren Suizidrate

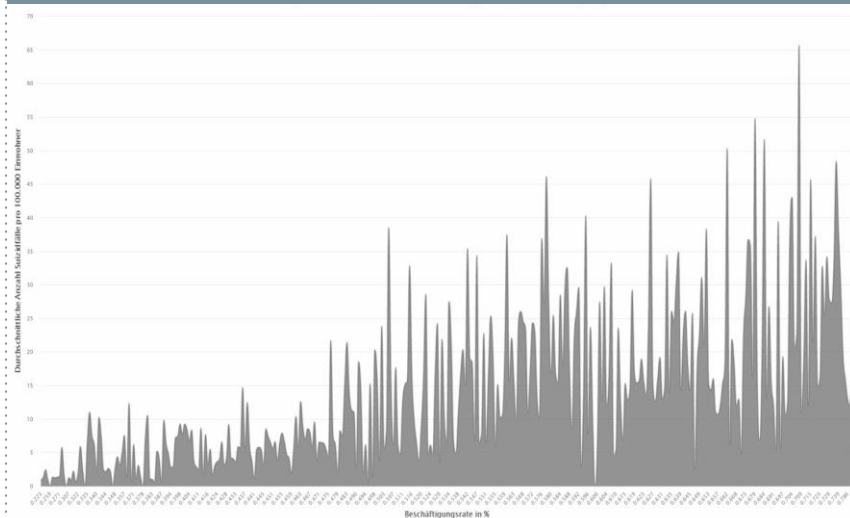
Suizidrate in Abh. zur Lebenserwartung



BESCHÄFTIGUNGSRATE - HARD WORK KILLS YOURSELF



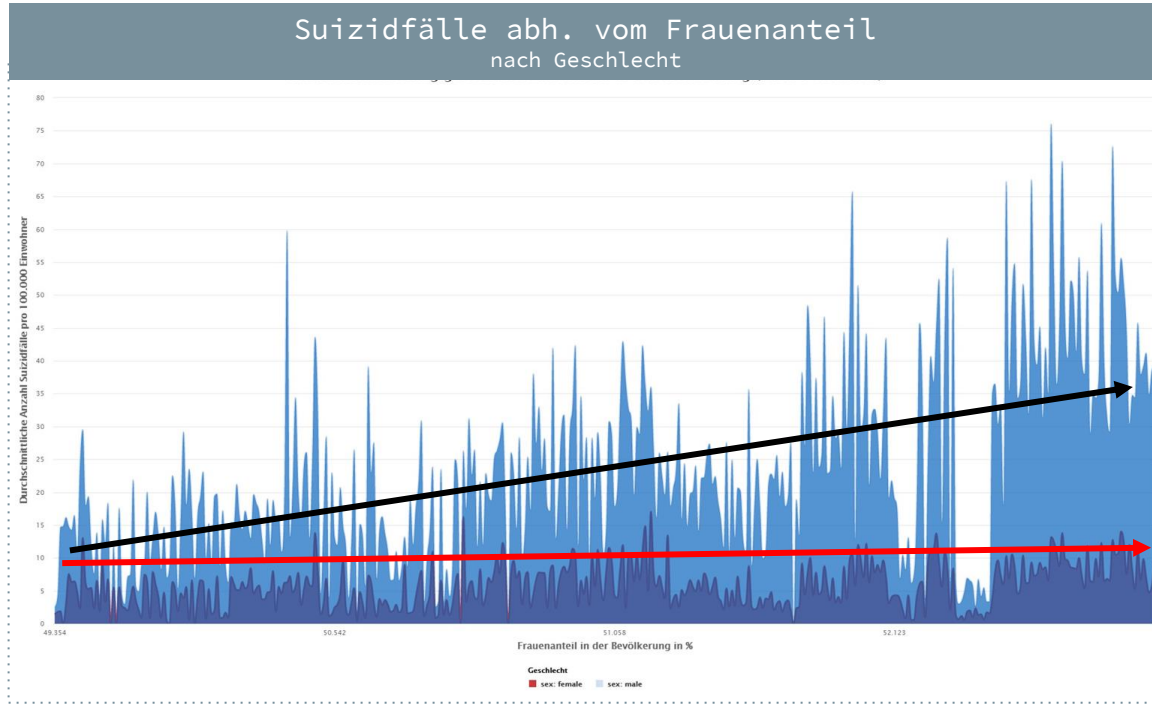
Suizidrate in Abh. zur Beschäftigungsrate



Suizidrate in Abh. zur Beschäftigungsrate
pro Geschlecht

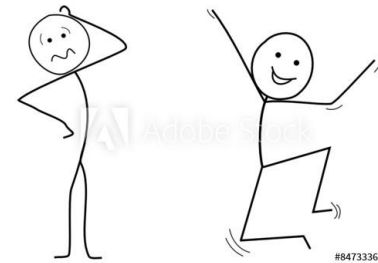


FRAUENANTEIL IN DER BEVÖLKERUNG (IN %)



INFO

- Erhöhter Frauenanteil in der Bevölkerung führt zu erhöhter Suizidrate bei Männern



#84733365

ZUSAMMENFASSUNG DATA EXPLORATION



LEICHTER/MITTLERER Zusammenhang zu Label

- Alter
- Geschlecht
- Frauenanteil
- Gesundheitsausgaben
- Lebenserwartung
- Beschäftigungsgrad



KEIN/WENIG Zusammenhang zu Label

Zu erwartender Zusammenhang
zum

- BIP
- BIP pro Kopf
- Einkommensungleichheit
(Gini-Coefficient)

nicht besonders stark



Erkenntnisse

- Kein erkennbarer Handlungsbedarf bei Ausreißern
- Zwei Attribute mit fehlenden Werten



DATA PREPARATION

- ANDREAS HERB, LARA MÜLLER -

DATA PREPARATION OVERVIEW

Basic Transformation

- Rollen
- Skalenniveaus
- Nominal/Binominale Attribute
- Duplicates (NA)

Feature Selection

- Multikollinearität
- Low Variance (NA)

Missing Values

- Simple Model:
Average
- Komplex Model:
Regression
- Exclude Values

Feature Scaling

- Standardisierung
- Normalisierung

Feature Engineering (NA)

Erstellen neuer Feature
in diesem Anwendungsfall
nicht notwendig

Outlier Handling (NA)

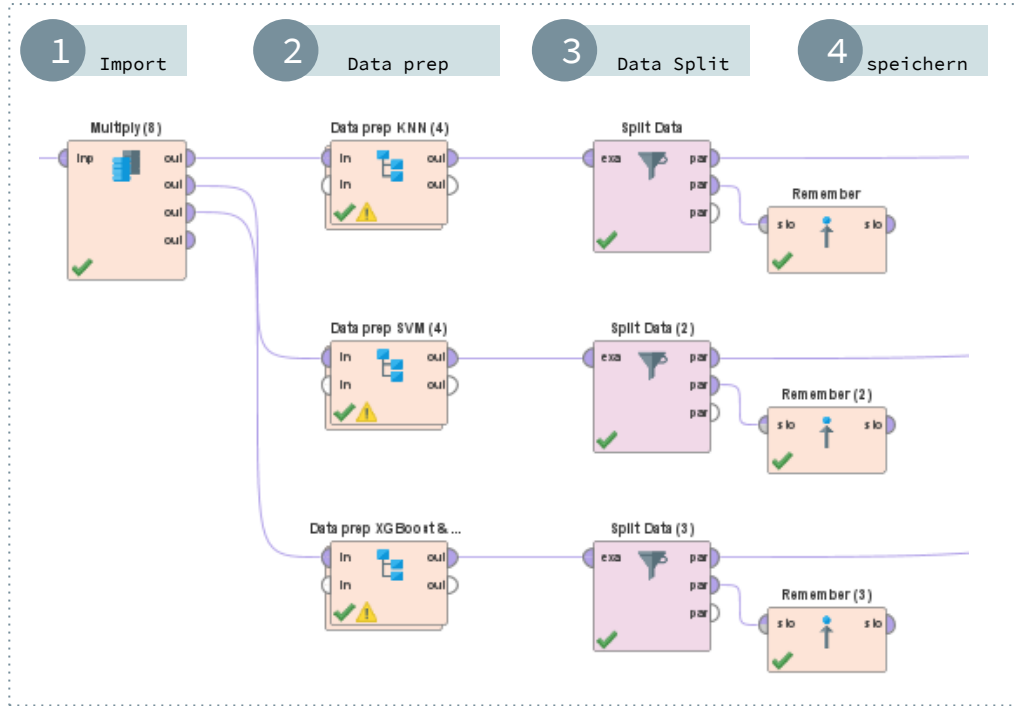
Keine zu behandelnden
Outlier in Datensatz

DATA PREPARATION - 3 DATA PREP PIPELINES (RECIPES)

Rapid Miner Process	Model / Preparation	General (roles, scales, dummy variables)	Missing Values	Feature Selection	Feature Scaling (Standardisierung)
2.1/ 3.1	KNN	X	X	X	X
2.2/ 3.2	SVM	X	X	X	SVM nutzt integriert scale Funktion, welche bessere Ergebnisse liefert
2.3/ 3.3	XGBoost / Random Forest	X	X	Tree Based Models nehmen Selektion der Feature selbst vor	Scaling bei Decision Tree Algorithmen nicht notwendig

siehe R Documentation "[Recommended Preprocessing](#)"

DATA PREPARATION GESAMTPROZESS - RAPID MINER



1. Datenset importieren
2. Data preparation je Algorithmus durchführen
3. Datenset mittels stratified sampling splitten
 - a. 25% Testdaten
 - b. 75% Trainingsdaten
4. Testdaten im Prozess speichern, um im späteren Verlauf darauf zugreifen zu können

DATA PREPARATION - MODULARER AUFBAU

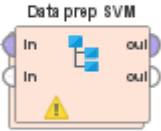
2

Data preparation

2.1



2.2

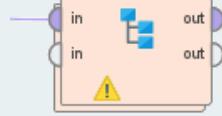


2.3



Modul 1

General Transforma...



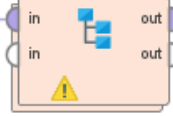
Entfernen nicht
relevanter
Attribute

Festlegen von
Rollen der
Attribute

Erstellen von Dummy
Variablen

Modul 2

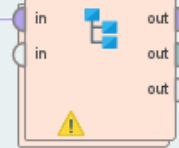
Missing Values



Auffüllen von
Missing Values
durch den "Impute
Missing Values"
Operator
-> Vorhersage
mittels XGBoost
Algorithmus

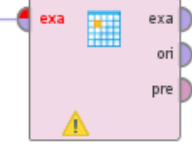
Modul 3

Feature Selection (2)



Modul 4

Normalize



Standardisierung
(Z-Transformation)
aller numerischen
Attribute mit
Ausnahme der dummy
Variablen

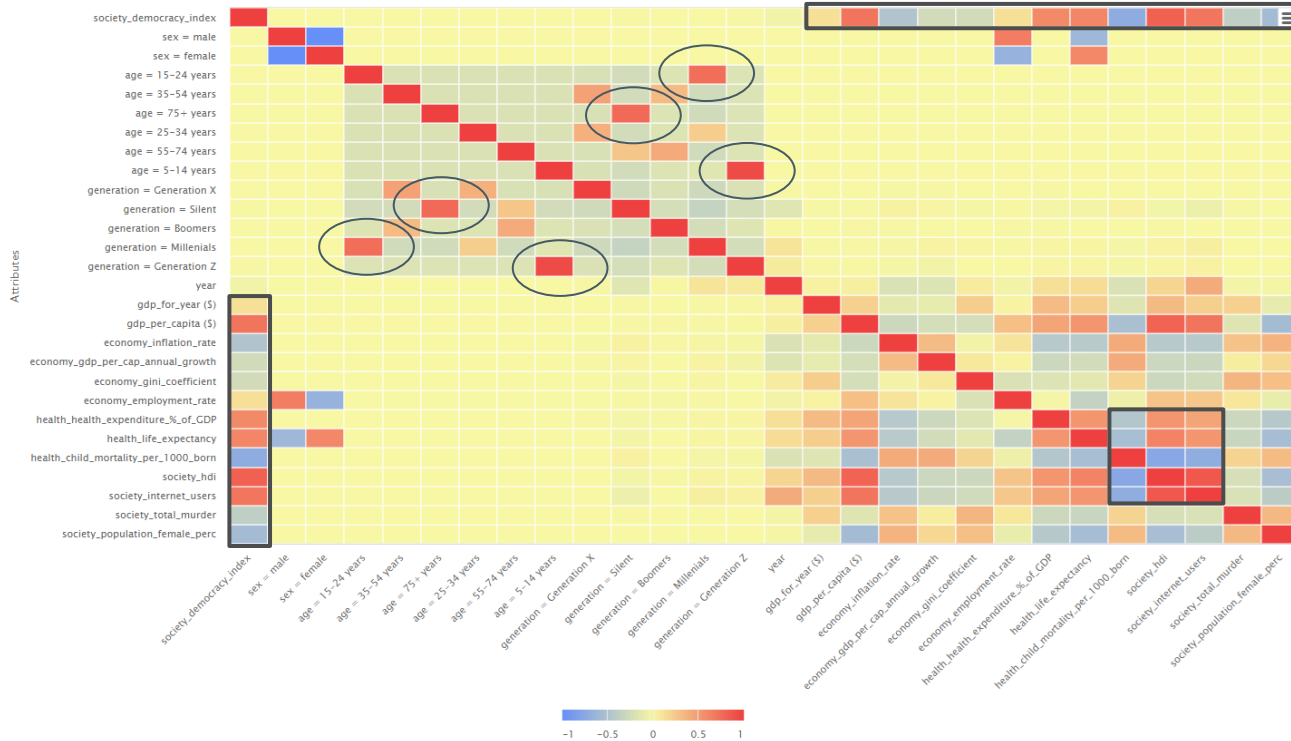
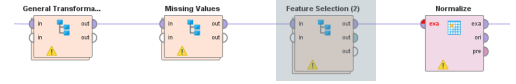
Data Prep KNN

Data Prep SVM

Data Prep XGBoost & Random Forest



ERGEBNIS NACH MISSING VALUES

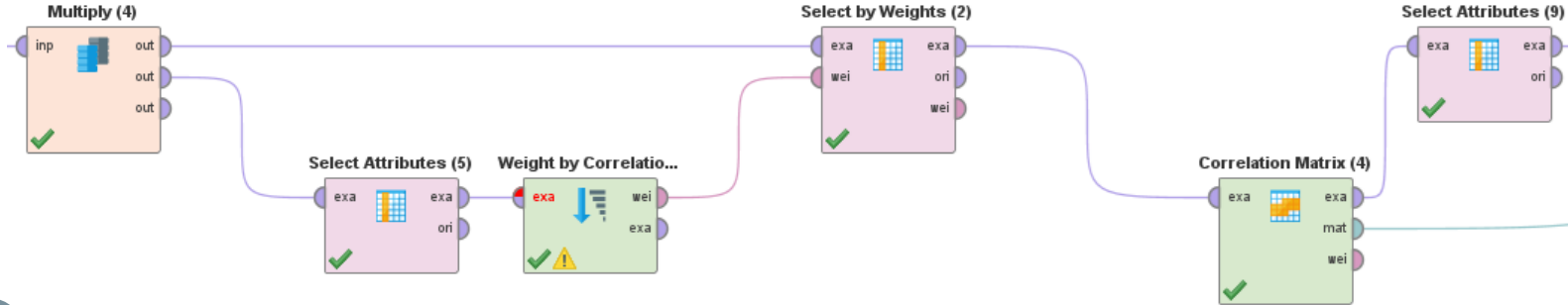
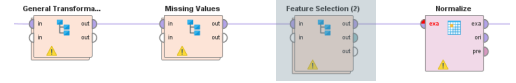


INFO

Auffällig hohe
Multikollinearität bei

- democracy_index
- gdp_per_capita
- hdi
- internet user
- child_mortality
- generation & age

MODUL 3 - FEATURE SELECTION



1

Sortieren der Attribute nach Stärke der Korrelation zum Label (siehe Folie 53)

Attribute mit einer normalisierten Gewichtung von $\leq 0,1$ werden über den "Select by Weights" Operator entfernt.

Da der Operator "Remove Correlated Attributes" für das Label relevante Feature entfernt, wurde entschieden den Selektionsprozess manuell durch Analyse durchzuführen.

weight relation	greater equals ▼
weight	0.1

2

Prüfen der Korrelationsmatrix auf Multikollinearität der übrigen Variablen (siehe Folie 54)

Anhand der Ergebnisse werden folgende Attribute über "Select Attributes" entfernt:

- Generation
- Health_Child_Mortality



MODELLING

- SEBASTIAN WÖLK, DAVID SCHALL -

MODELLING OVERVIEW

Choose Algorithms

- K-Nearest Neighbor
- Support Vector Machine
- Random Forest
- Gradient Boosted Tree

Hyperparametertuning

- Auswahl geeigneter Parameter für jeden Algorithmus
- Definition eines Wertebereichs

Cross Validation

- Festlegen der Anzahl Folds
- Festlegen der Splitting-Methode

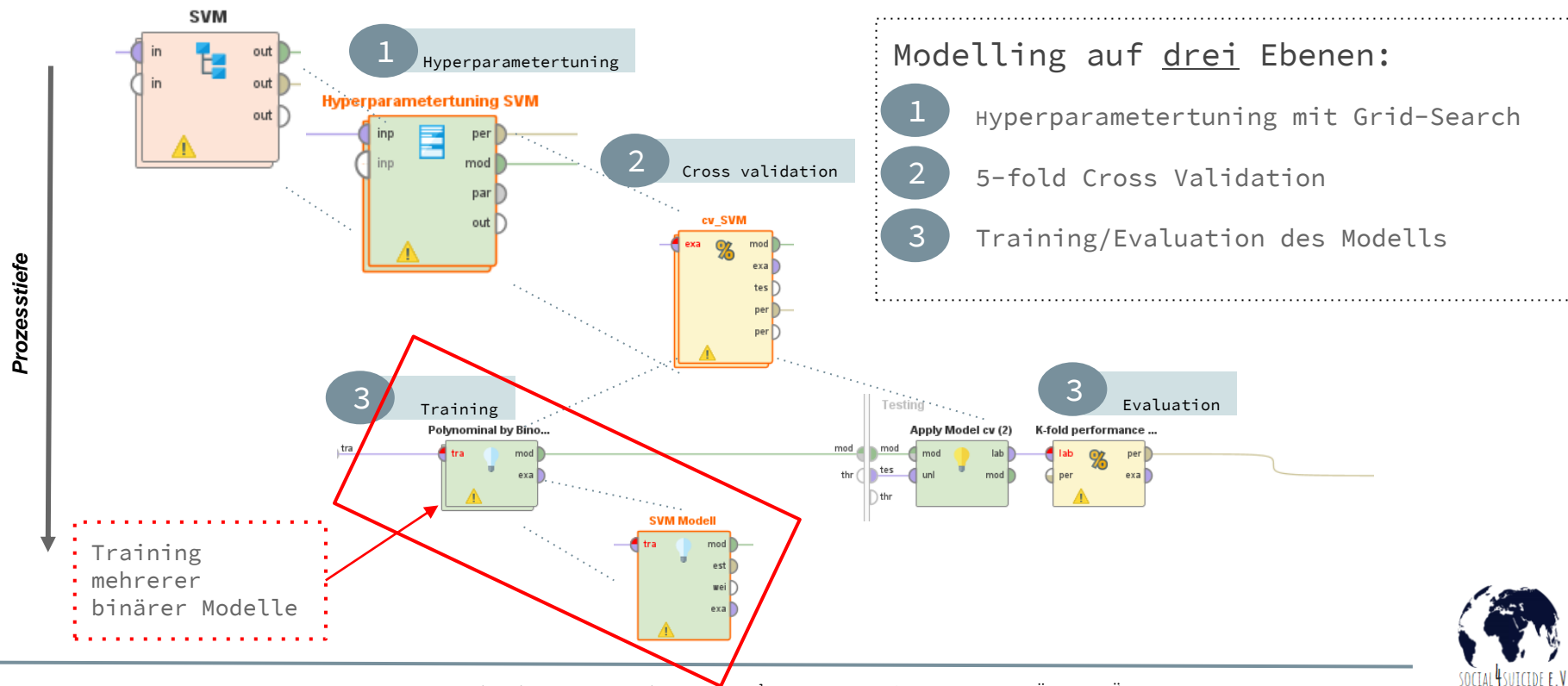
Training

- Trainieren auf algorithmus-spezifischen Trainingsdatensätzen
- Eventuelle Anpassung der Data Preparation

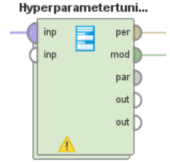
Testing & Performance

- Definition der Performanceparameter
- Messen der Performance auf den Validierungsdaten der cross validation

SUPPORT VECTOR MACHINE: PROZESSEBENEN



SUPPORT VECTOR MACHINE: HYPERPARAMETER TUNING



Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators

- cv_SVM (Cross Validation)
- Polynomial by Binominal Classification (SVM Modell (Support Vector Machine))
- Apply Model cv (2) (Apply Model)
- K-fold performance SVM (Performance (CI))

Parameters

Selected Parameters

- SVM Modell.kernel_gamma
- SVM Modell.C

Grid/Range

Min	Max	Steps	Scale
0.0	0.3	6	linear

Kernel Gamma:

Min: 0.0

Max: 0.3

Steps: 6

Scale: Linear

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators

- cv_SVM (Cross Validation)
- Polynomial by Binominal Classification (SVM Modell (Support Vector Machine))
- Apply Model cv (2) (Apply Model)
- K-fold performance SVM (Performance (CI))

Parameters

Selected Parameters

- SVM Modell.kernel_gamma
- SVM Modell.C

Grid/Range

Min	Max	Steps	Scale
-1.0	15	5	linear

C:

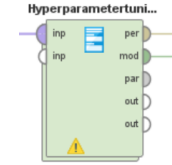
Min: -1.0

Max: 15

Steps: 5

Scale: Linear

RANDOM FOREST: HYPERPARAMETER TUNING



Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators

- Cross Validation (Cross Validation)
- Random Forest (Random Forest)
- Attribute weights RF (2) (Weights to Data)
- Aufsteig. Sortierung RF (Sort)
- Attribute weights RF (3) (Remember)
- Apply Model cv (Apply Model)
- K-fold performance RF (Performance (Cia

Parameters

Selected Parameters

- Random Forest.number_of_trees
- Random Forest.maximal_depth

Grid/Range

Min	Max	Steps	Scale
50	80	5	linear



Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators

- Cross Validation (Cross Validation)
- Random Forest (Random Forest)
- Attribute weights RF (2) (Weights to Data)
- Aufsteig. Sortierung RF (Sort)
- Attribute weights RF (3) (Remember)
- Apply Model cv (Apply Model)
- K-fold performance RF (Performance (Cia

Parameters

Selected Parameters

- Random Forest.number_of_trees
- Random Forest.maximal_depth

Grid/Range

Min	Max	Steps	Scale
40	80	7	linear

Anzahl der Bäume:

Min: 50

Max: 80

Steps: 5

Scale: Linear

Max. Tiefe der Bäume:

Min: 40

Max: 80

Steps: 7

Scale: Linear

K-NEAREST NEIGHBOR: HYPERPARAMETER TUNING



Operators

- cv KNN (3) (Cross Validation)
- k-NN (3) (k-NN)
- Apply Model (5) (Apply Model)
- K-fold performance KNN (3) (Performan

Parameters

Selected Parameters

- k-NN (3).k
- k-NN (3).weighted_vote

Grid/Range

Min	Max	Steps	Scale
1	21	21	linear



Operators

- cv KNN (3) (Cross Validation)
- k-NN (3) (k-NN)
- Apply Model (5) (Apply Model)
- K-fold performance KNN (3) (Performan

Parameters

Selected Parameters

- k-NN (3).k
- k-NN (3).weighted_vote

Grid/Range

Min	Max	Steps	Scale
1	21	21	linear

Value List

☒ true
☐ false

k-Neighbors:

Min: 1

Max: 21

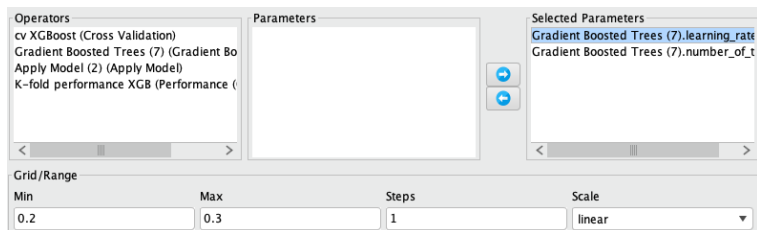
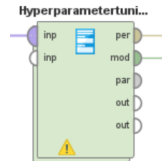
Steps: 21

Scale: Linear

weighted_vote:

True

GRADIENT BOOSTED TREE: HYPERPARAMETER TUNING



learning rate
min: 0.2
max: 0.3
Steps: 1

Anzahl der Bäume:
Min: 130
Max: 160
Steps: 1
Scale: Linear

TRAININGSDATEN: PERFORMANCEÜBERSICHT

Phase	Algorithmus	Accuracy	Recall	Precision
Training	Support Vector Machine	80.05%	80.09%	80.88%
	K-Nearest Neighbor	79.88 %	79.92%	79.97%
	Random Forest	82.00%	82.02%	82.38%
	Gradient Boosted Tree	83.98%	84.02%	84.03%



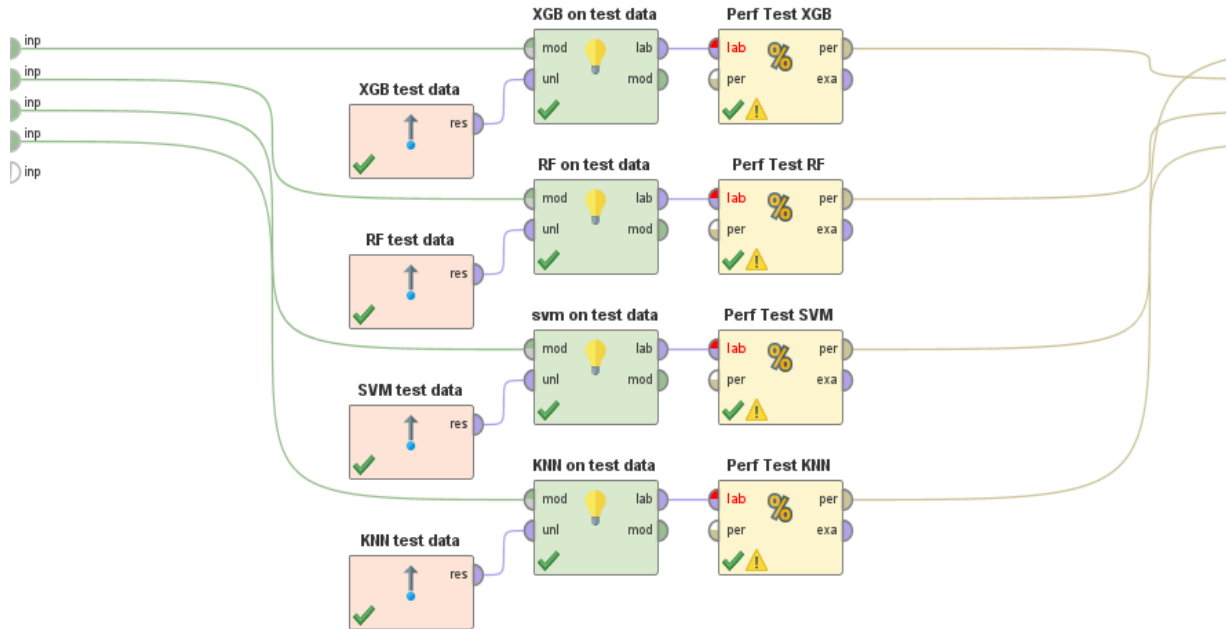
MODEL EVALUATION

- SEBASTIAN WÖLK, ANDREAS HERB -

GESAMTPROZESS: EVALUATION

1 Modelle 2 Testdaten 3 Modelling & Performance 4 Output

Evaluation



INFO

- Anwendung des trainierten Modells auf die Testdaten
- Performancemessung
 - Accuracy
 - Precision
 - Recall

EVALUATION: PERFORMANCEÜBERSICHT

Phase	Algorithmus	Accuracy	Recall	Precision
Training	Support Vector Machine	80.05%	80.09%	80.88%
	K-Nearest Neighbor	79.88 %	79.92%	79.97%
	Random Forest	82.00%	82.02%	82.38%
	Gradient Boosted Tree	83.98%	84.02%	84.03%
Testing	Support Vector Machine	78.62%	78.65%	79.40%
	K-Nearest Neighbor	79.31%	79.35%	79.38%
	Random Forest	81.03%	81.07%	81.61%
	Gradient Boosted Tree	82.76%	82.80%	82.93%

EVALUATION: GRADIENT BOOSTED TREE (XG BOOST)

TRAIN

accuracy: 83.98% +/- 1.40% (micro average: 83.98%)

	true no_risk	true medium_risk	true low_risk	true high_risk	class precision
pred. no_risk	778	6	84	2	89.43%
pred. medium_risk	7	680	95	69	79.91%
pred. low_risk	89	103	689	9	77.42%
pred. high_risk	1	85	8	779	89.23%
class recall	88.91%	77.80%	78.65%	90.69%	

Accuracy: 83.98%

Recall: 84.02%

Precision: 84.03%

TEST

accuracy: 82.76%

	true no_risk	true medium_risk	true low_risk	true high_risk	class precision
pred. no_risk	244	1	36	0	86.83%
pred. medium_risk	5	234	35	22	79.05%
pred. low_risk	41	36	220	2	73.58%
pred. high_risk	1	20	1	262	92.25%
class recall	83.85%	80.41%	75.34%	91.61%	

Accuracy: 82.76%

Recall: 82.80%

Precision: 82.93%

EVALUATION: MODELLAUSWAHL

GRADIENT BOOSTED TREE



Gute Performance auf Trainings- und Testdaten



Gute Interpretierbarkeit von Decision Tree-Modellen



Transparente Modellbildung



Sehr bekannter und performanter Klassifikator



Geringe Anforderungen an Data Preparation



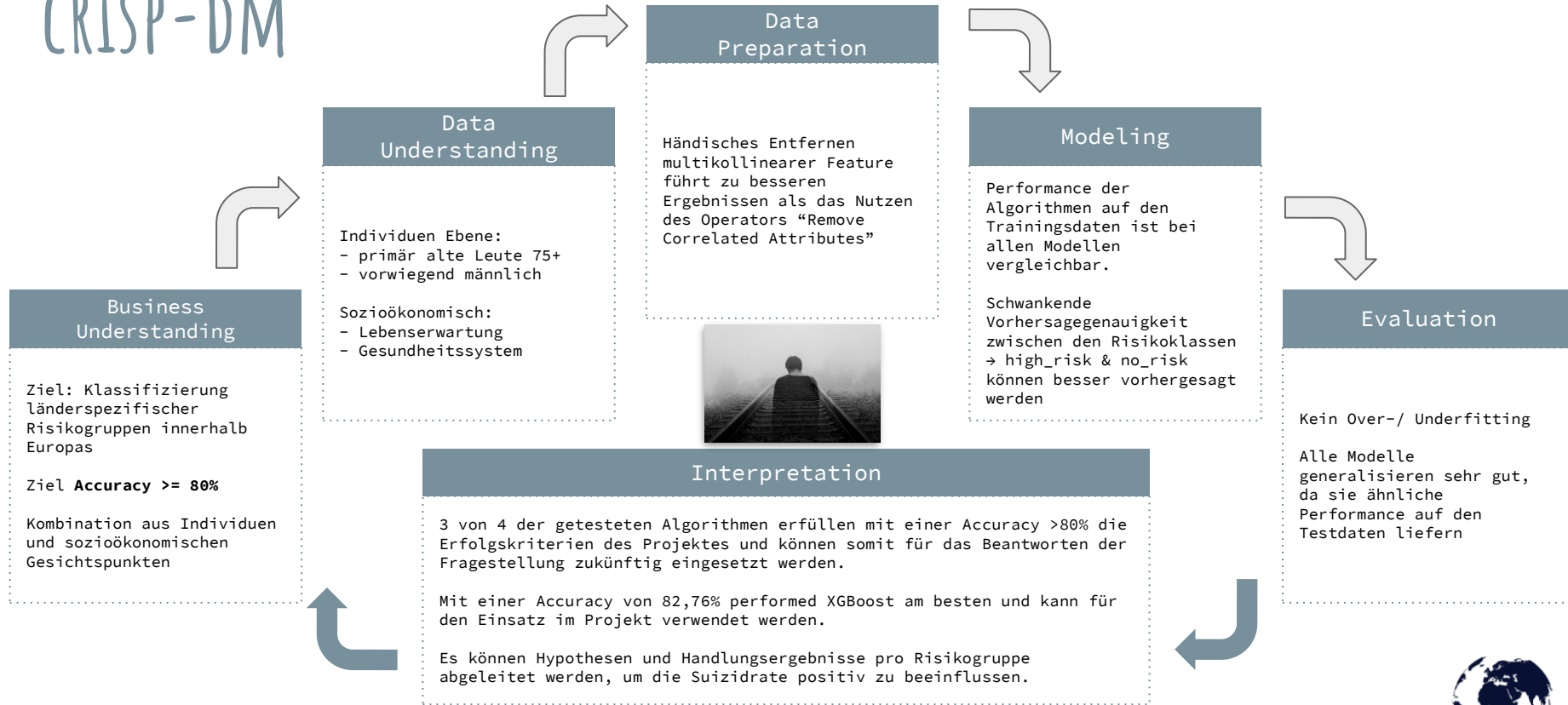
Eventuell rechenintensiv bei großen Datenmengen



RESULT PRESENTATION & INTERPRETATION


- DAVID SCHALL, LARA MÜLLER -

CRISP-DM



EINORDNEN DER RISIKOKLASSEN IN DEN GESAMTKONTEXT

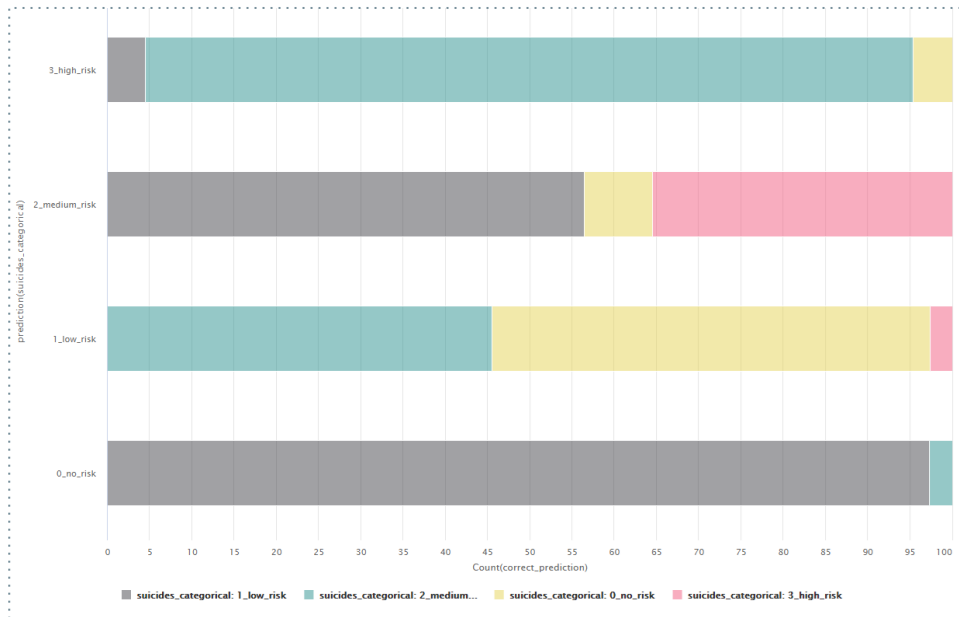
Risikoklasse	Abgeleitete Maßnahmen
high_risk	<ul style="list-style-type: none">• Bereitstellen von Finanzpaketen für das Länderbudget• Initiieren und Durchführung weiterer Studien auf Risikogruppenebene• Erarbeiten eines Maßnahmenplans zur Suizidprävention
medium_risk	<ul style="list-style-type: none">• Erarbeitung eines individuellen Maßnahmenplans• Finanzielle Unterstützung dedizierter sozialer Projekte im Land
low_risk	<ul style="list-style-type: none">• Beobachtung des Risikostatus und ermittelter Einflussfaktoren• Ggf. Verteilen von Informationen zum Thema Suizidprävention
no_risk	<ul style="list-style-type: none">• Land wird als nicht risikoreich eingestuft → keine Aktionen notwendig



Mit Hilfe der Risikoklassen wird Höhe und Ausmaß der Unterstützung definiert.

Da mit der Unterstützung von **high_risk** Risikogruppen der Einsatz hoher finanzieller Mittel verbunden ist, ist die genaue Vorhersage dieser Risikoklasse von höchster Bedeutung.

ANALYSE FALSCH-KLASSIFIZIERUNG



Bei Falsch-Klassifizierungen wird meist die benachbarte Risikoklasse vorhergesagt

high_risk

high_risk kann sehr zuverlässig vorhergesagt werden. Im Falle einer Falsch-Klassifikation besteht maximal das Risiko einer medium_risk Klassifizierung

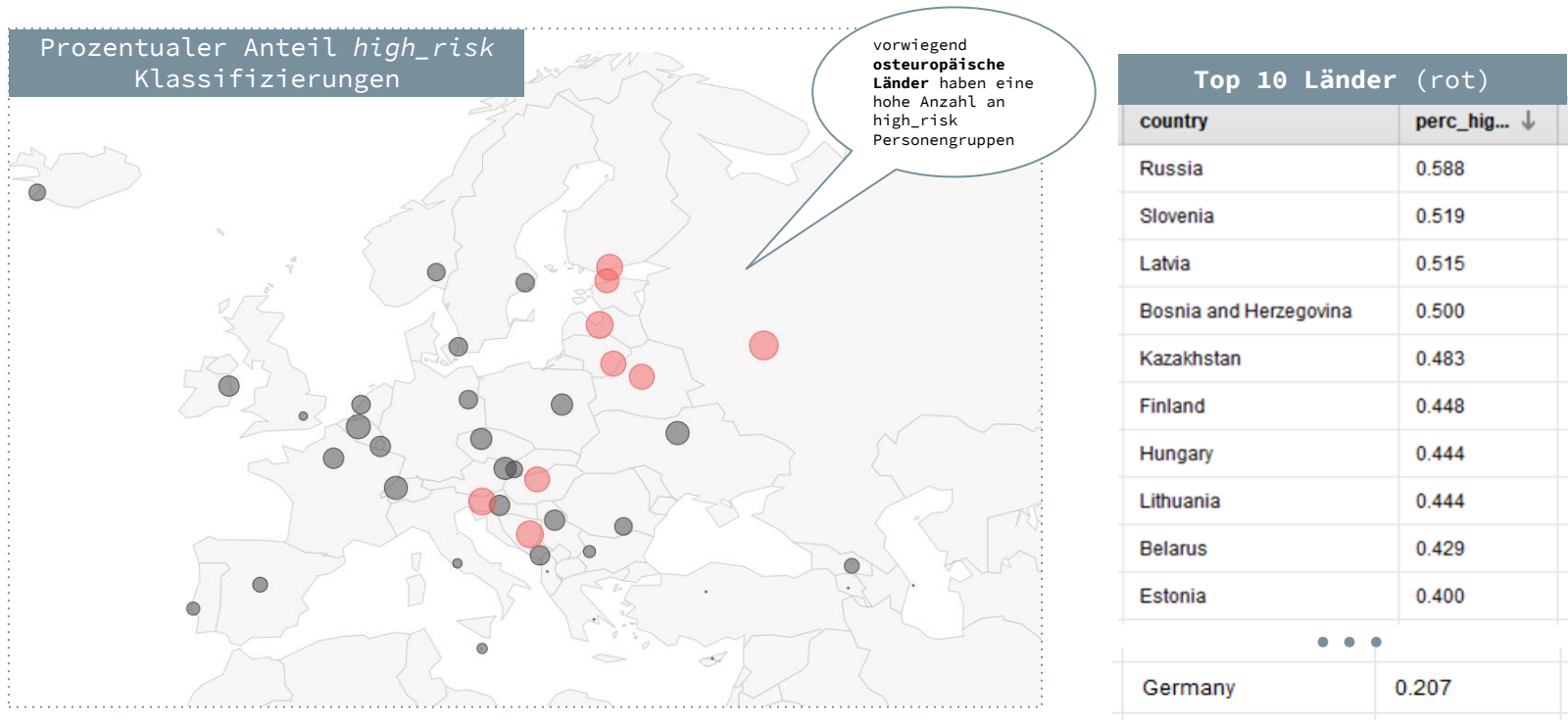
medium_risk & low risk

medium_risk & low_risk bedarf näherer Analysen
→ Hohes Risiko einer Fehlinvestition bei da Personengruppe tatsächlich darunterliegender Klasse angehören
→ Hohes Risiko für mangelnden Einsatz, da Personengruppe tatsächlich höher klassifiziert werden müssten

no_risk

Falsch-Klassifikation bei no_risk bleiben ohne größere Konsequenz, da meist die nächsthöhere Klasse low_risk gewählt wird (ebenfalls keine/kaum Aktionen)

HIGH-RISK KLASSIFIKATION TESTDATEN - XGBOOST



RESULT INTERPRETATION: ZUSAMMENFASSUNG

Methodik



high_risk & no_risk Personengruppen können mit einer hohen Vorhersagegenauigkeit zur Entscheidung über Investitionen herangezogen werden



Zwischenklassen (low_risk & medium_risk) können durch Falsch-Klassifizierungen zu Fehlinvestitionen führen und bedürfen einer genaueren Analyse

Optimierung: Zusammenführen beider Klasse zu einer

Domain



Finanzielle Investitionen in das Gesundheitssystem und gesunde Beschäftigungsverhältnisse (Work-Life-Balance & gleichberechtigte Bezahlung) können die Suizidrate positiv beeinflussen



insbesondere das zunehmende Alter und das Geschlecht (männlich) sind ausschlaggebend für hohe Suizidraten

→ Beratungsangebote & Kampagnen für ausgewählte Risikogruppen pro Land unterstützen einen effizienten Einsatz der Ressourcen

Ausblick



Risikoklassifizierung als Priorisierung weiterer Analysen auf Individuen-Ebene



Prüfen der Anwendbarkeit des Modells auf andere Kontinente (z.B. Asien, Amerika)



Steigern der Vorhersagegenauigkeit der Algorithmen durch Hinzufügen weiterer Feature

Vielen Dank für
eure
Aufmerksamkeit
;-)



KLASSIFIZIERUNG DES SUIZIDRISIKOS VON PERSONENGRUPPEN IN LÄNDERN EUROPAS

- ANDREAS HERB, DAVID SCHALL, SEBASTIAN WÖLK, LARA MÜLLER -