Machine Learning Project

Ein Vergleich von ML-Modellen zur Klassifikation von Drogenkonsum

Luis Rastetter, Salih Kelmendi

Agenda

01 Business Understanding

02 Data Understanding

03 Data Preparation

04 Modeling

05 Evaluierung

Business Understanding

Literatur und Zielsetzung

Business Understanding

Mithilfe von Persönlichkeitsdaten und soziodemografischen Informationen können Machine Learning Modelle trainiert werden, um:

1. Vorherzusagen, **welche** Personen in Zukunft **bestimmte** Drogen konsumieren

2. Zu analysieren, **welche Merkmale besonders relevant** für den Konsum legaler oder illegaler Substanzen sind

Ergebnisse der Literaturrecherche

Big Five Personality Traits and Illicit Drug Use: Specificity in Trait-Drug Associations

- Bestimmte Persönlichkeitsmerkmale hängen mit Drogenkonsum zusammen
- Niedrige Verträglichkeit/Gewissenhaftigkeit -> Hoher Drogenkonsum

Big Five personality traits and alcohol, nicotine, cannabis, and gambling disorder comorbidity

- Alkoholkonsum assoziiert mit: Neurotizismus
- Keine Assoziation mit: Extraversion, Offenheit, Verträglichkeit

Ursachen für Drogenkonsum

- Psychische, soziale und biologische Faktoren, Genetische Veranlagung

Bisherige Predictions

- F1-Score, Accuracy als Metriken genutzt
- Log. Regression, Ridge Classifier, Support Vector Machines, Random Forest Classifier als Modelle

SMART Ziele

Spezifisch

Ein Vergleich verschiedener Machine-Learning-Modelle soll durchgeführt werden, um vorherzusagen, welche Art von legalen oder illegalen Drogen Personen zukünftig konsumieren könnten – basierend auf Persönlichkeitsmerkmalen und Umfragedaten.

Messbar

Der Erfolg der Modelle wird anhand konkreter Kennzahlen wie Trainings- und Testfehler, F1-Score und Accuracy gemessen.

Erreichbar

Die Durchführung des Vergleichs ist im Rahmen eines Projektzeitraums mit vorhandenen Daten (aus dem Paper) und gängigen ML-Methoden realistisch durchführbar.

Relevant

Das Ziel knüpft direkt an das zuvor analysierte Paper an und hat einen klaren Bezug zur Studien- und Forschungspraxis, insbesondere zur Anwendung von Data Science im Bereich Verhaltensprognosen.

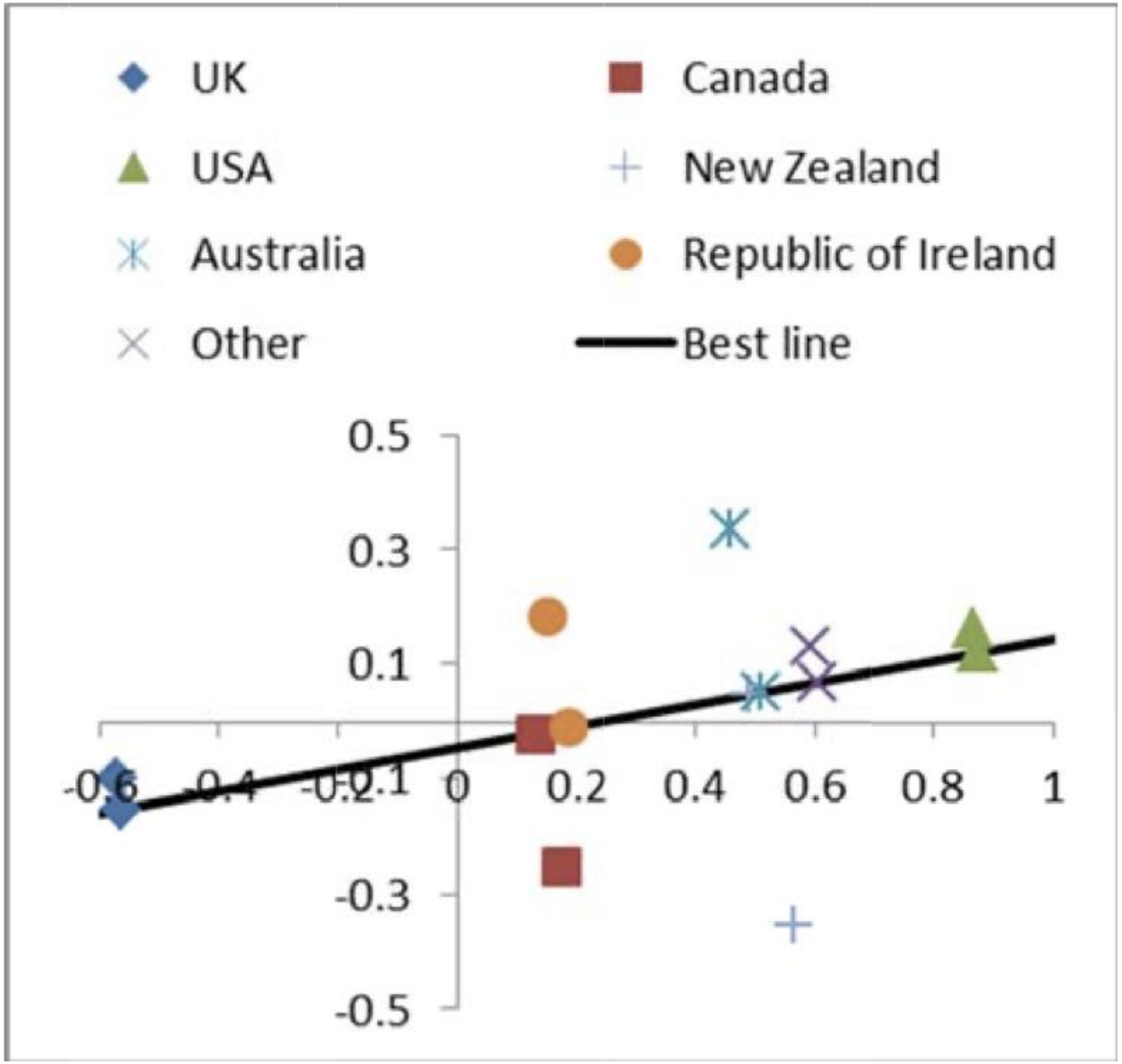
Zeitgebunden

Der Vergleich soll bis zum Abschluss des 4. Semesters abgeschlossen sein.

Data Understanding

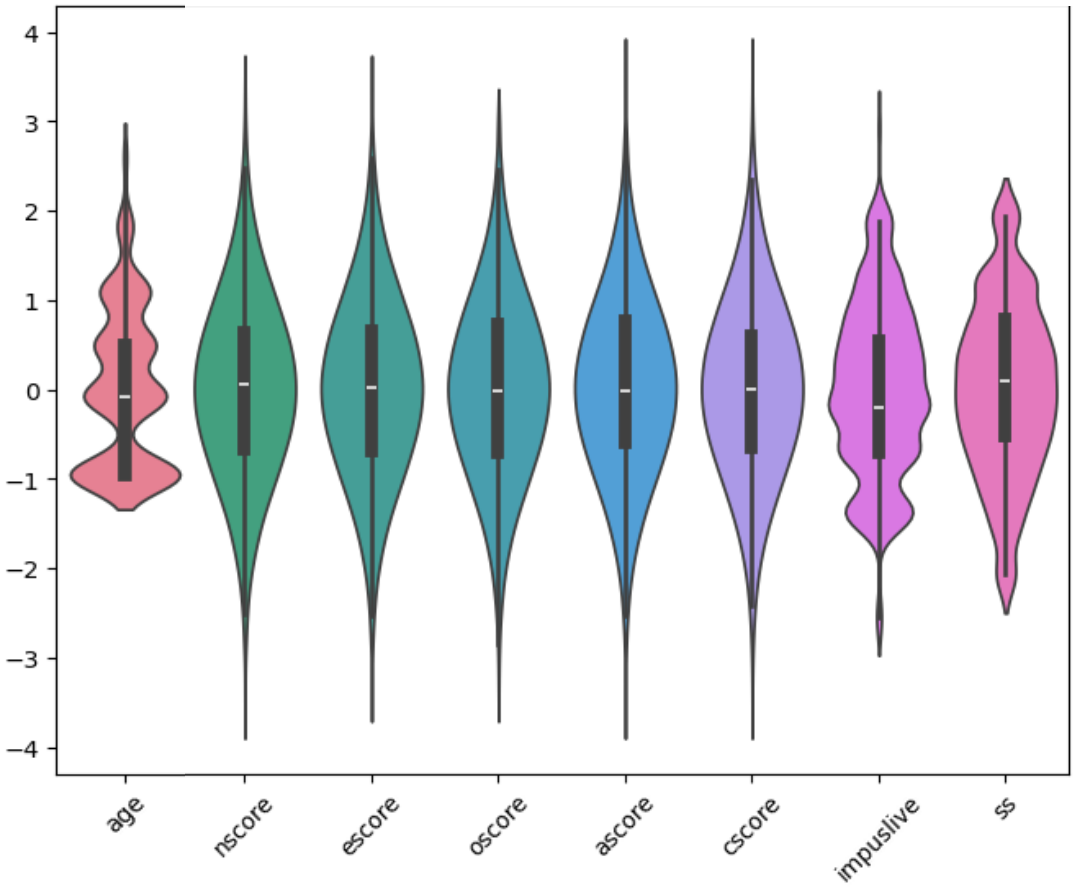
Der Datensatz

- Online Umfrage von Elaine Fehrman zwischen 2011 und 2012.
- 12 Features
- 1885 valide Teilnehmende aus UK, USA, Canada, Neuseeland, Irland und Australien
- Geschlechterverteilung: 943 Männer / 942 Frauen.
- Ordinal/Nominal feature quanitification wurde angewendet


UK
Canada
USA
New Zealand
Australia
Republic of Ireland
Other
Best line
0.5
0.3
0.1
-0.3
-0.5
-0.6
-0.4
-0.2
-0.1
0
0.2
0.4
0.6
0.8
1

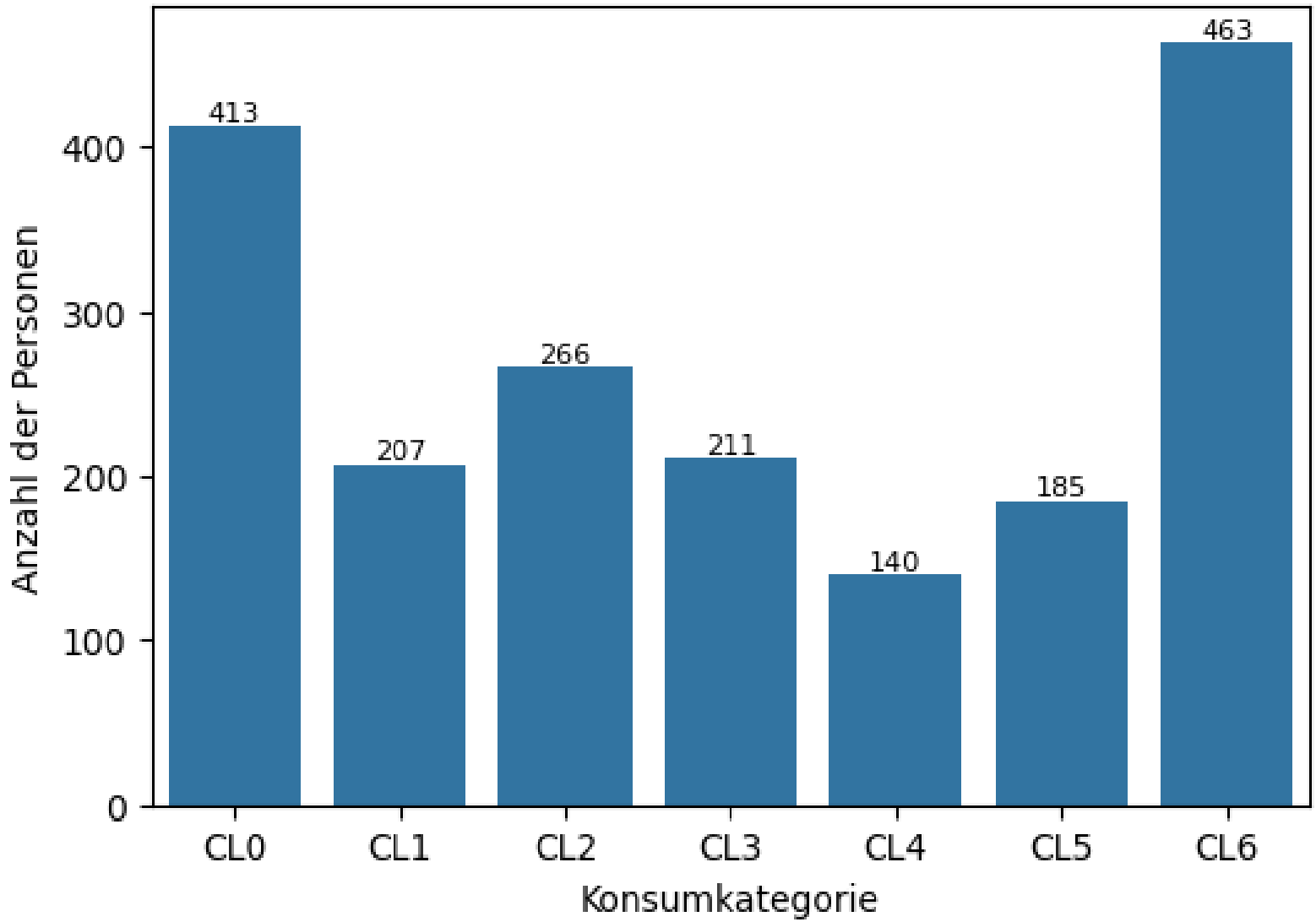
Spalte1	nscore	escore	oscore	ascore	cscore	impuslive	ss
count	1885	1885	1885	1885	1885	1885	1885
mean	0,000047	-0,00016	-0,00053	-0,00025	-0,00039	0,007216	-0,00329
std	0,998106	0,997448	0,996229	0,99744	0,997523	0,954435	0,963701
min	-3,46436	-3,27393	-3,27393	-3,46436	-3,46436	-2,90161	-2,07848
25%	-0,67825	-0,69509	-0,71727	-0,60633	-0,65253	-0,71126	-0,52593
50%	0,04257	0,00332	-0,01928	-0,01729	-0,00665	-0,21712	0,07987
75%	0,62967	0,63779	0,7233	0,76096	0,58489	0,52975	0,7654
max	3,27393	3,27393	2,90161	3,46436	3,46436	2,90161	1,92173

Violinplot

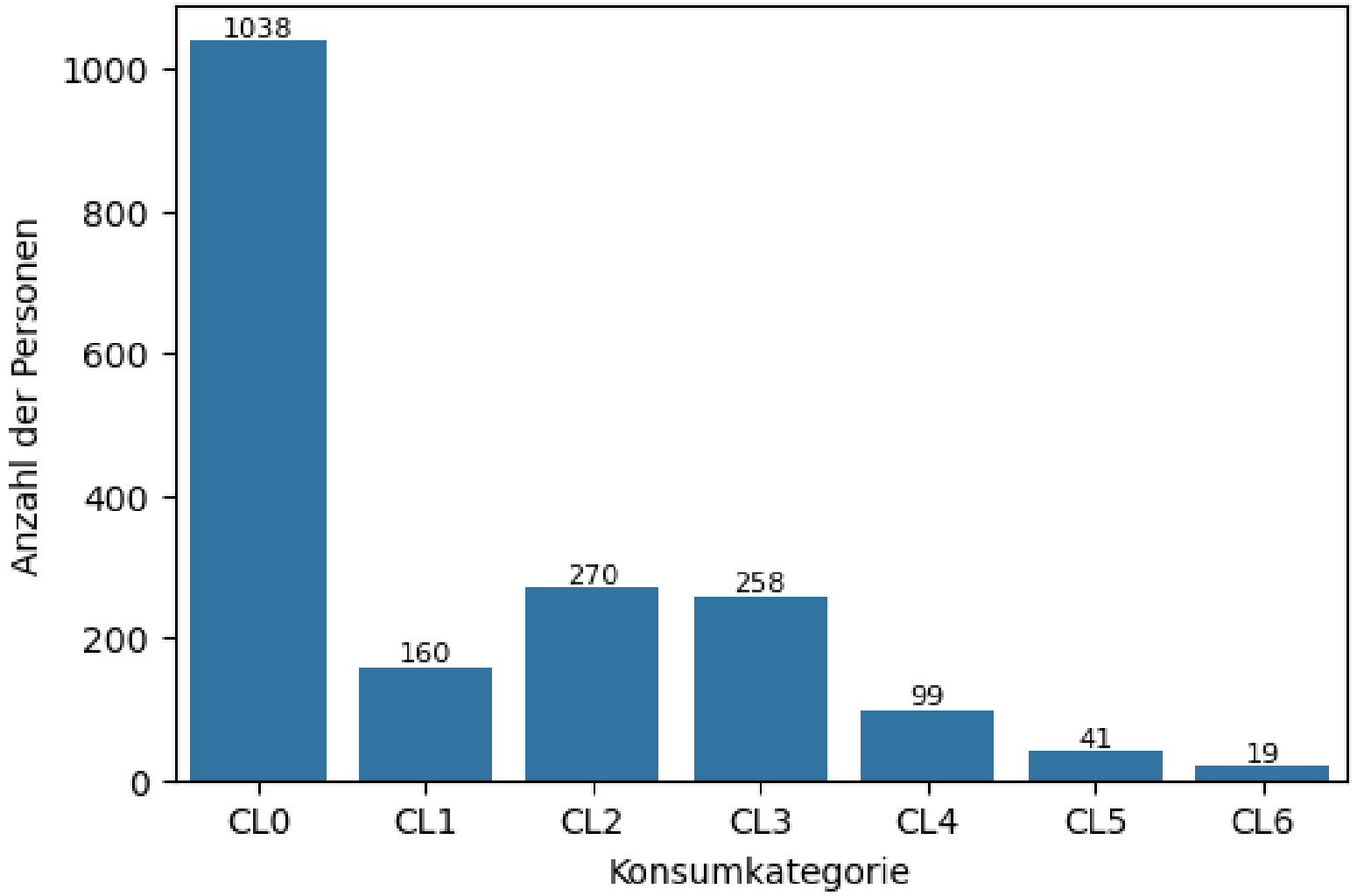


- nscore–cscore: weitgehend normalverteilt, symmetrisch
- impulsive: leicht linksschief, Ausreißer nach unten
- ss: symmetrisch, breite Verteilung

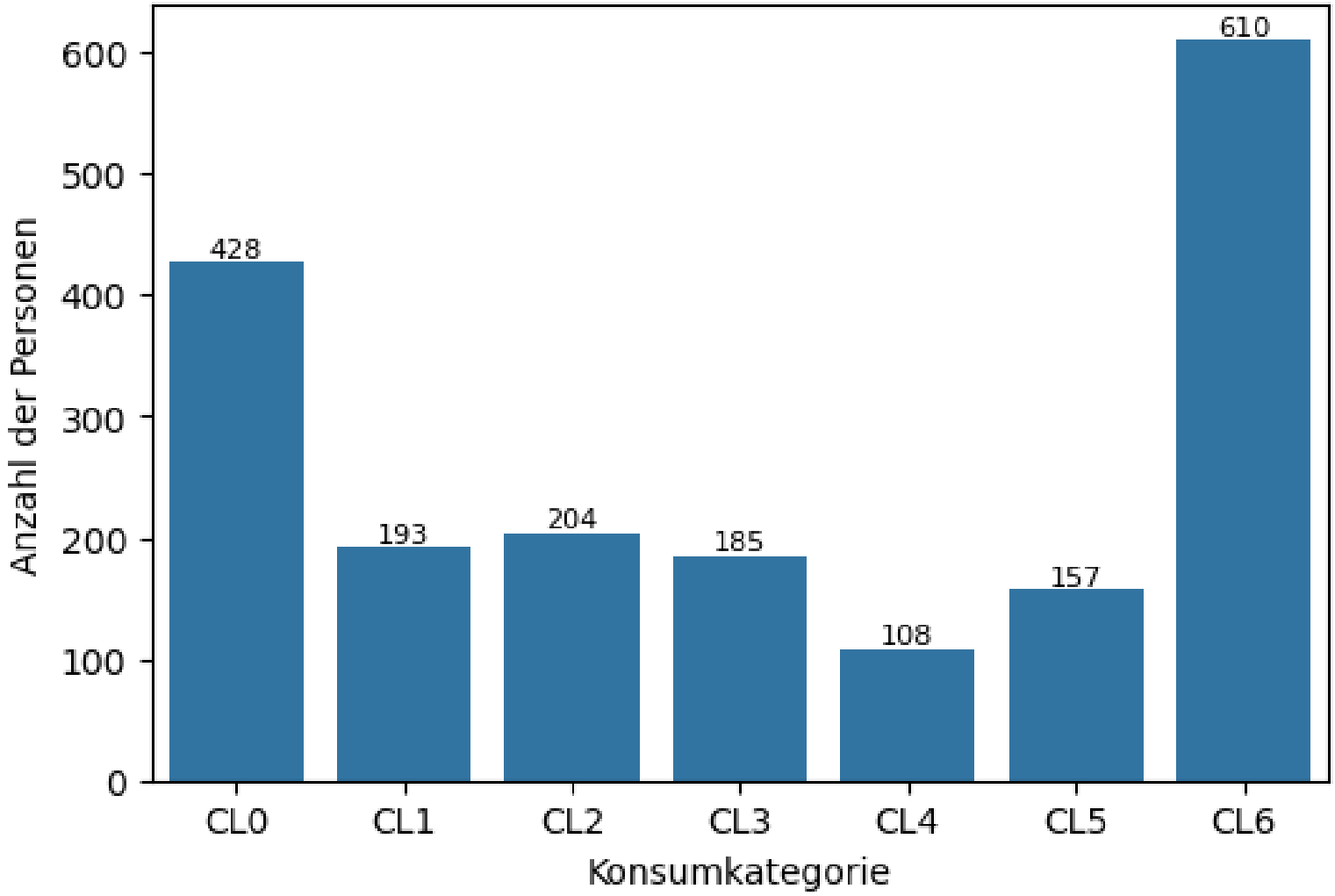
Verteilung der Konsumkategorien für Cannabis



Verteilung der Konsumkategorien für Kokain



Verteilung der Konsumkategorien für Nikotin



Data Preparation

Anpassen des Datensatzes

- Überprüfen auf Duplikate

```
print("Anzahl Duplikate im Datensatz: ", sum(X.duplicated()))

if sum(X.duplicated()) == 0:
    print("Keine Duplikate im Datensatz.")
```

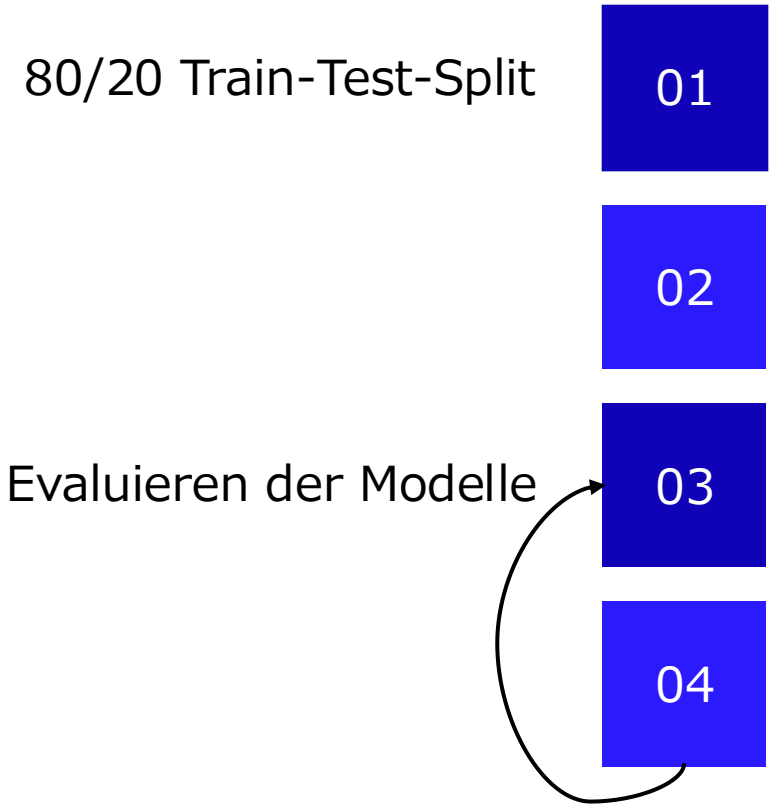
  - Keine Duplikate im Datensatz

- Entfernen von unwichtiger Features

  - Gender und Ethnicity wurden entfernt

Modeling

Ablauf während des Modellierens



Trainieren der Modelle

- **Decision Tree**, KNN, SVM, **Log. Regression**, Random Forest

Optimieren der Modelle

Hyperparameteroptimierung mittels Grid Search

Optimierung des DT – Kurzer Einblick

```python
from sklearn.model_selection import GridSearchCV
dt_models = {}
y_pred_dt = {}
best_params_dt = {}
for drug in drug_names:
    print(f"Grid Search für {drug} läuft...")
    class_weights = compute_class_weight('balanced', classes=np.unique(y_train[drug]), y=y_train[drug])
    class_weight_dict = dict(zip(np.unique(y_train[drug]), class_weights))
    param_grid = {
        'criterion': ['gini', 'entropy'],
        'max_depth': [5, 10, 15] if drug in ['coke', 'nicotine'] else [3, 5, 7],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 4]
    }
    base_model = DecisionTreeClassifier(
        random_state=42,
        class_weight=class_weight_dict
    )
    grid_search = GridSearchCV(
        estimator=base_model,
        param_grid=param_grid,
        cv=3,
        scoring='f1_macro',
        n_jobs=-1
    )
    grid_search.fit(X_train, y_train[drug])
    best_model = grid_search.best_estimator_

    dt_models[drug] = best_model
    y_pred_dt[drug] = best_model.predict(X_test)
    best_params_dt[drug] = grid_search.best_params_
    print(f"Beste Parameter für {drug}: {grid_search.best_params_}")
```

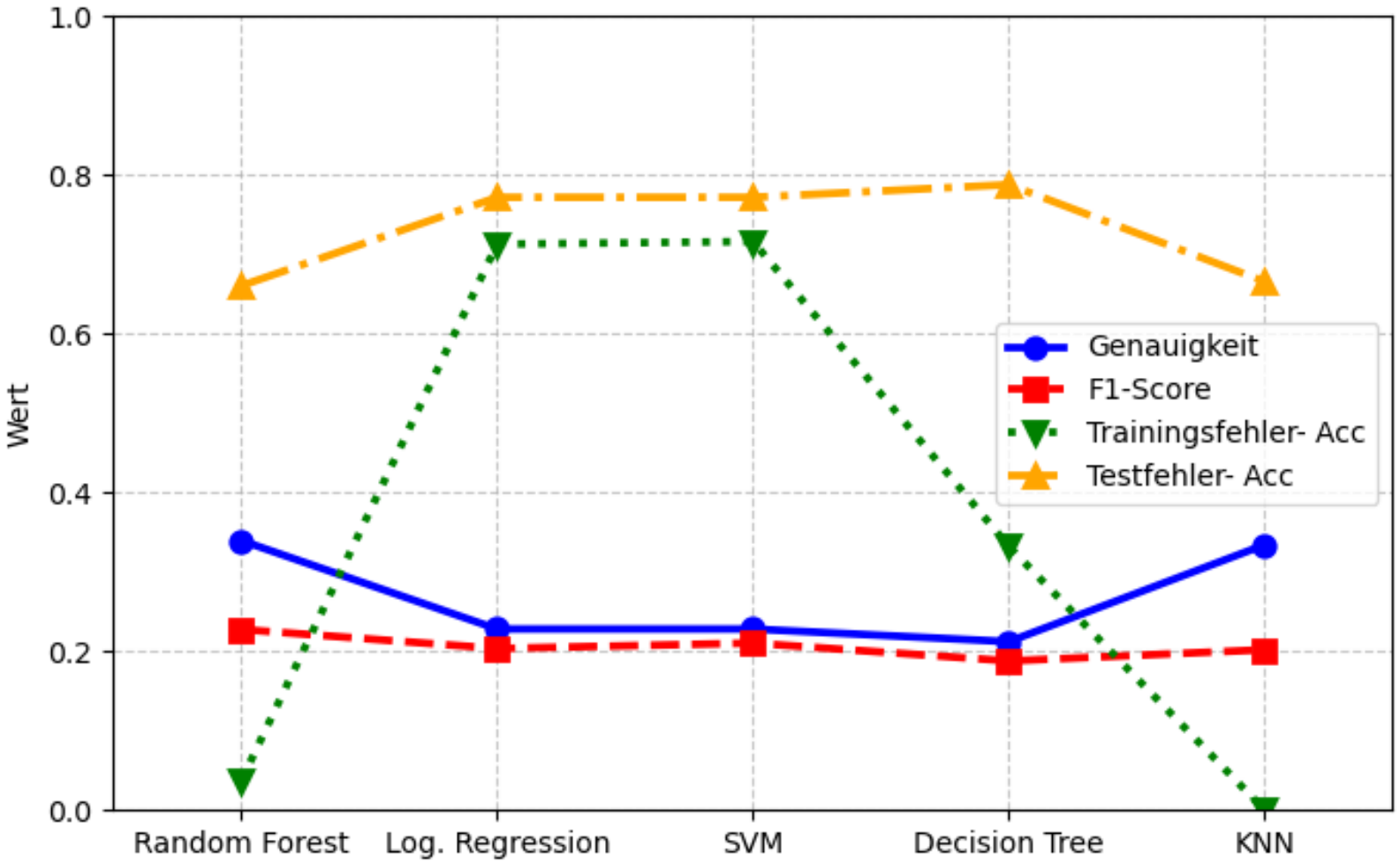
```python
dt_models = {}
y_pred_dt = {}

for drug in drug_names:
    class_weights = compute_class_weight('balanced', classes=np.unique(y_train[drug]), y=y_train[drug])
    class_weight_dict = dict(zip(np.unique(y_train[drug]), class_weights))

    dt_model = DecisionTreeClassifier(
        random_state=42,
        class_weight=class_weight_dict,
        criterion='entropy',
        max_depth=10 if drug in ['coke', 'nicotine'] else 5,
        min_samples_leaf=1,
        min_samples_split=2
    )

    dt_model.fit(X_train, y_train[drug])
    dt_models[drug] = dt_model
    y_pred_dt[drug] = dt_model.predict(X_test)
```

Modellbewertung für Nikotin



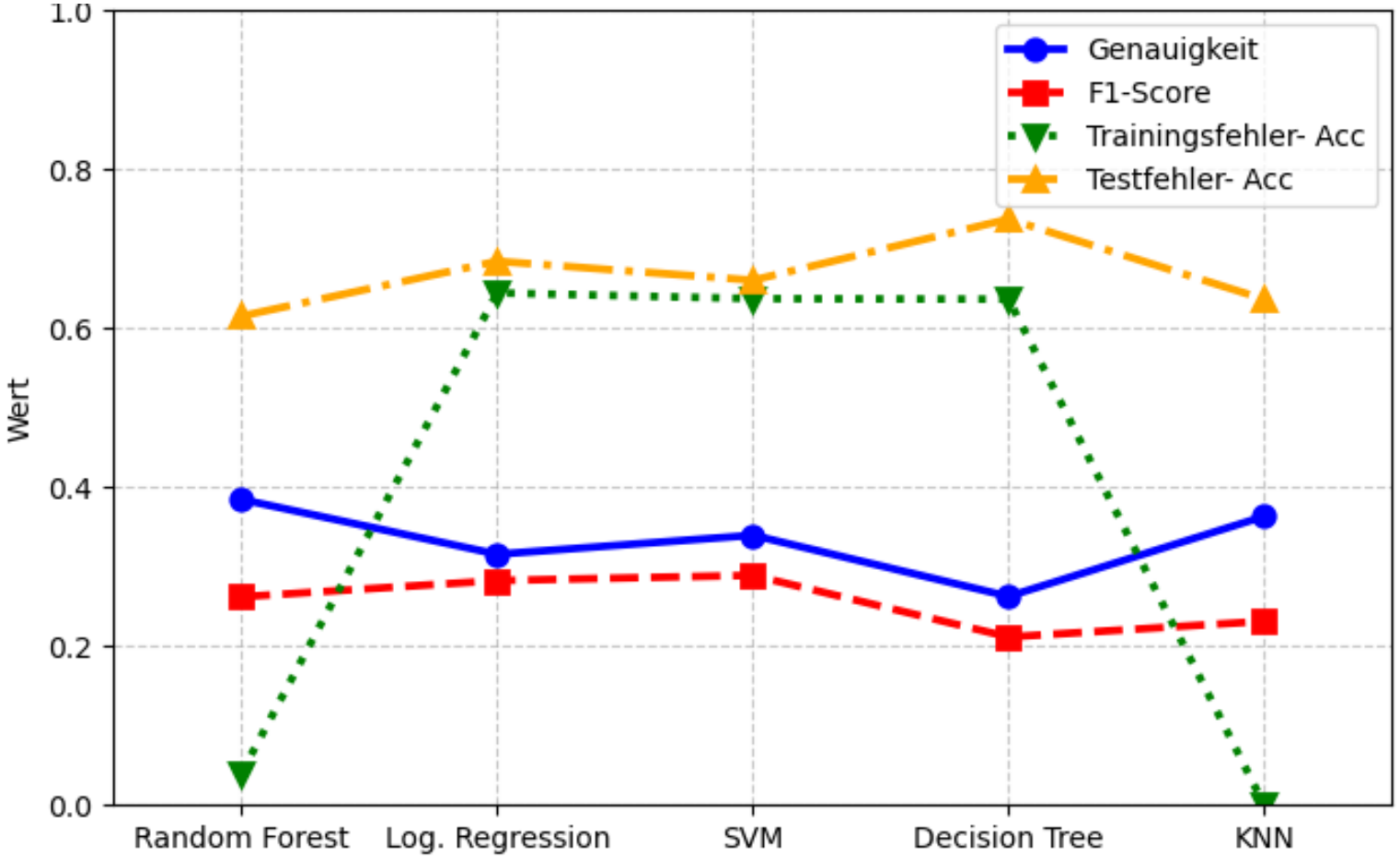
Class	Precision	Recall	F1-Score	Support
CL0	0,37	0,23	0,28	91
CL1	0,2	0,47	0,28	34
CL2	0,22	0,2	0,21	44
CL3	0,07	0,09	0,08	33
CL4	0,03	0,05	0,04	19
CL5	0,14	0,29	0,19	34
CL6	0,36	0,16	0,23	122
accuracy		0,21		377
macro avg	0,2	0,22	0,19	377
weighted avg	0,27	0,21	0,22	377

Decision Tree

Class	Precision	Recall	F1-Score	Support
CL0	0,39	0,26	0,32	91
CL1	0,19	0,44	0,27	34
CL2	0,11	0,09	0,1	44
CL3	0,19	0,36	0,25	33
CL4	0,08	0,26	0,13	19
CL5	0,09	0,09	0,09	34
CL6	0,5	0,19	0,27	122
accuracy		0,23		377
macro avg	0,22	0,24	0,2	377
weighted avg	0,32	0,23	0,24	377

Log. Regression

Modellbewertung für Cannabis



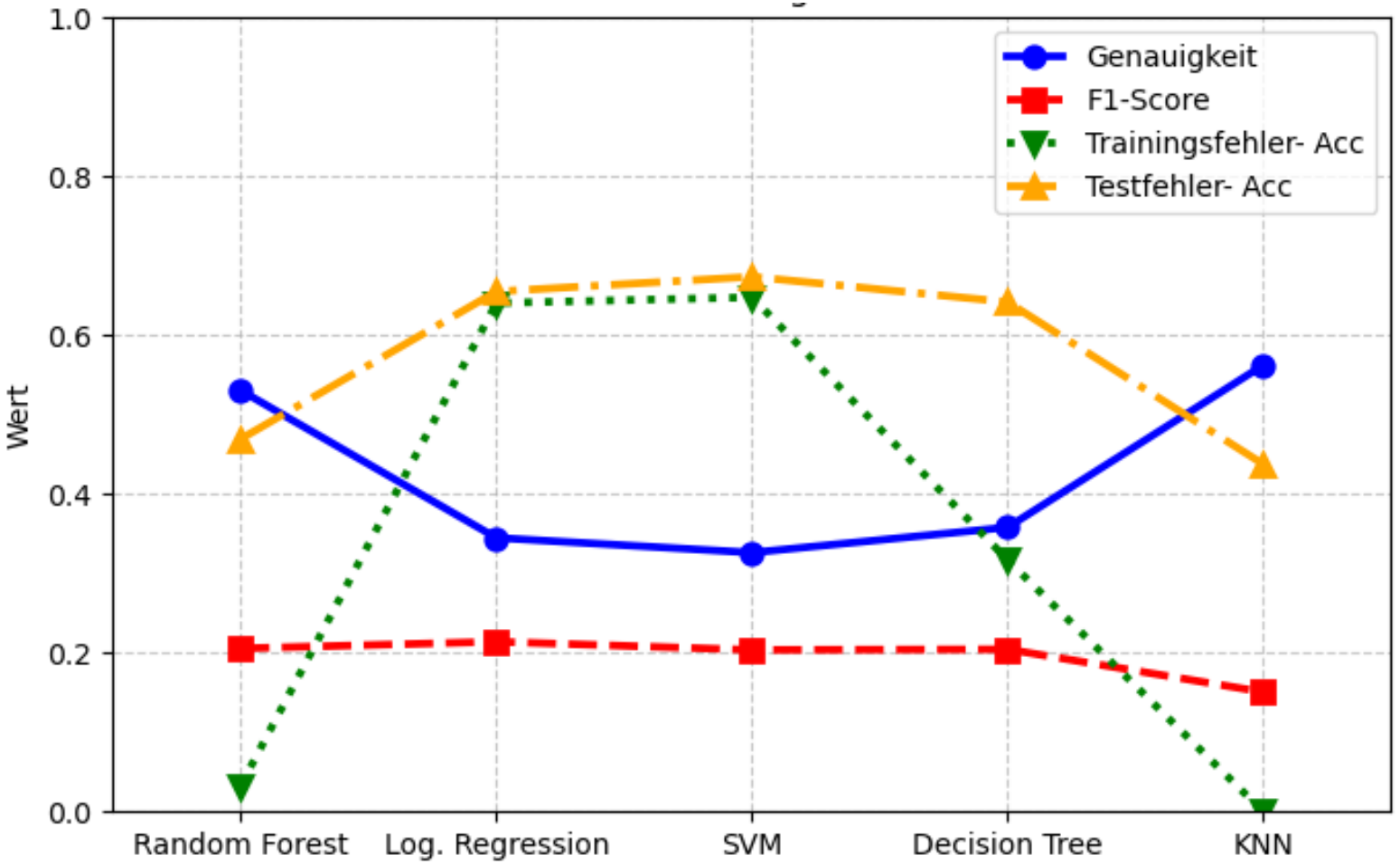
Class	Precision	Recall	F1-Score	Support
CL0	0,35	0,22	0,27	86
CL1	0,15	0,5	0,23	28
CL2	0,3	0,13	0,18	55
CL3	0,21	0,25	0,22	57
CL4	0,04	0,04	0,04	27
CL5	0,07	0,06	0,07	32
CL6	0,48	0,46	0,47	92
accuracy		0,26		377
macro avg	0,23	0,24	0,21	377
weighted avg	0,29	0,26	0,26	377

Decision Tree

Class	Precision	Recall	F1-Score	Support
CL0	0,56	0,45	0,5	86
CL1	0,23	0,57	0,32	28
CL2	0,25	0,25	0,25	55
CL3	0,3	0,18	0,22	57
CL4	0,09	0,11	0,1	27
CL5	0,18	0,28	0,22	32
CL6	0,46	0,3	0,37	92
accuracy		0,32		377
macro avg	0,29	0,31	0,28	377
weighted avg	0,36	0,32	0,32	377

Log. Regression

Modellbewertung für Kokain



Class	Precision	Recall	F1-Score	Support
CL0	0,71	0,44	0,54	224
CL1	0,28	0,43	0,34	30
CL2	0,18	0,25	0,21	44
CL3	0,17	0,21	0,19	43
CL4	0,07	0,12	0,09	25
CL5	0,04	0,12	0,06	8
CL6	0	0	0	3
accuracy		0,36		377
macro avg	0,21	0,23	0,2	377
weighted avg	0,49	0,36	0,4	377

Decision Tree

Class	Precision	Recall	F1-Score	Support
CL0	0,77	0,43	0,55	224
CL1	0,19	0,43	0,26	30
CL2	0,13	0,09	0,11	44
CL3	0,14	0,14	0,14	43
CL4	0,2	0,24	0,22	25
CL5	0,08	0,38	0,13	8
CL6	0,05	0,67	0,09	3
accuracy		0,34		377
macro avg	0,22	0,34	0,21	377
weighted avg	0,52	0,34	0,39	377

Log. Regression

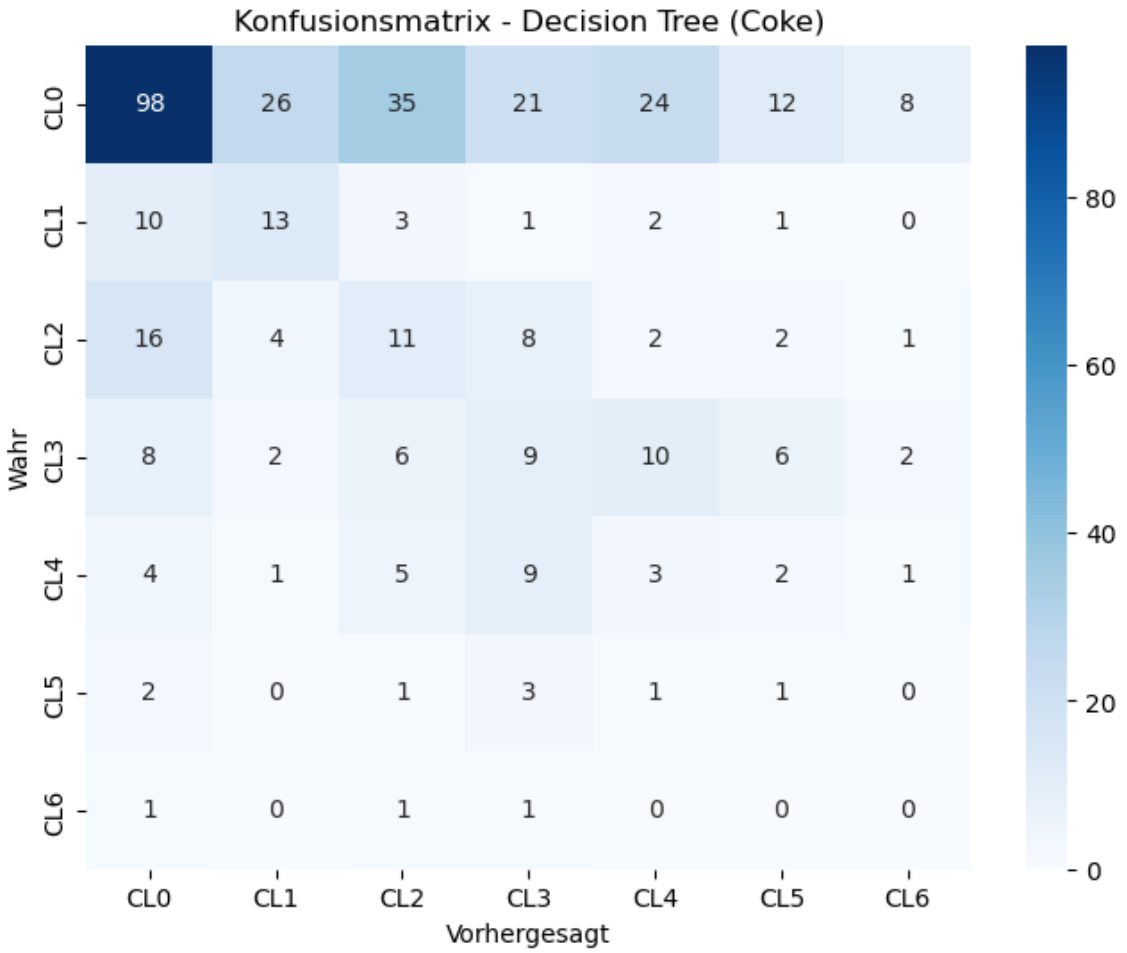
Evaluierung

Wo sind die Schwächen der Modelle?

Schwachstellen unserer Modelle

Am Beispiel Decision Tree - Kokain

- In den Testdaten deutlich mehr CL0 als CL6
- Multiklassenklassifikation
- Drogenkonsum ist sehr komplex und schwer vorhersagbar
- Merkmalsgewichtung hat wenig Aussagekraft



Auswahl des Modells

Grundsätzlich: Log. Regression dem Decision Tree vorzuziehen

- Geringerer Testfehler

  → bessere Generalisierungsleistung

- Stabilere Genauigkeit und F1-Score

  → robuster bei unausgeglichenen Klassen

- Weniger anfällig für Overfitting

  → Trainings- und Testfehler sind ähnlich

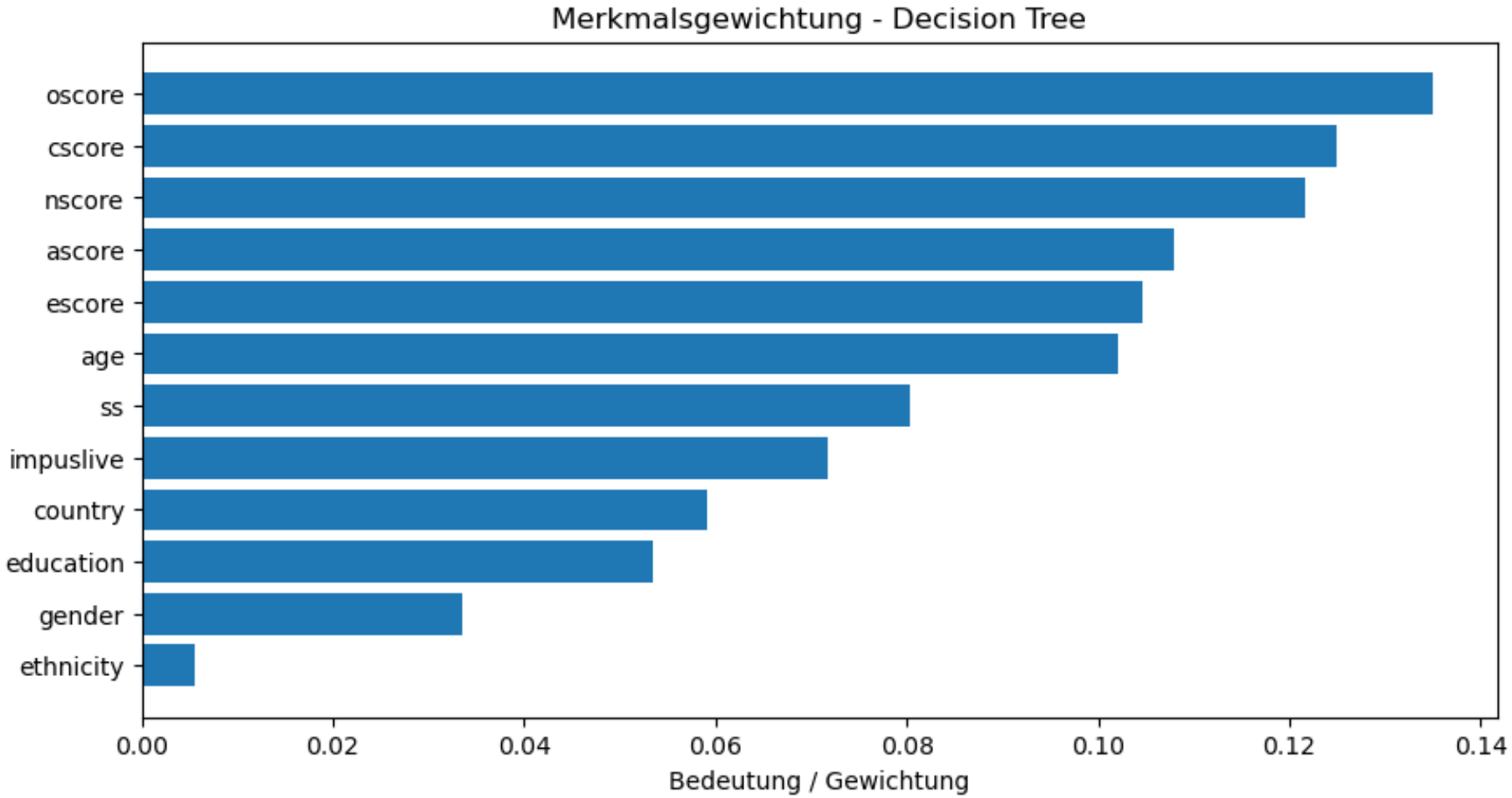
- Einfacheres Modell

  → leichter interpretierbar und schneller trainierbar

Jedoch...

Merkmalsgewichtung

Am Beispiel Decision Tree - Kokain



Fazit

Ausblick

Welche Änderungen können vorgenommen werden, um das Ergebnis zu verbessern?

Grundsätzlich

- Multiklassifikation zu einer binären Klassifikation machen

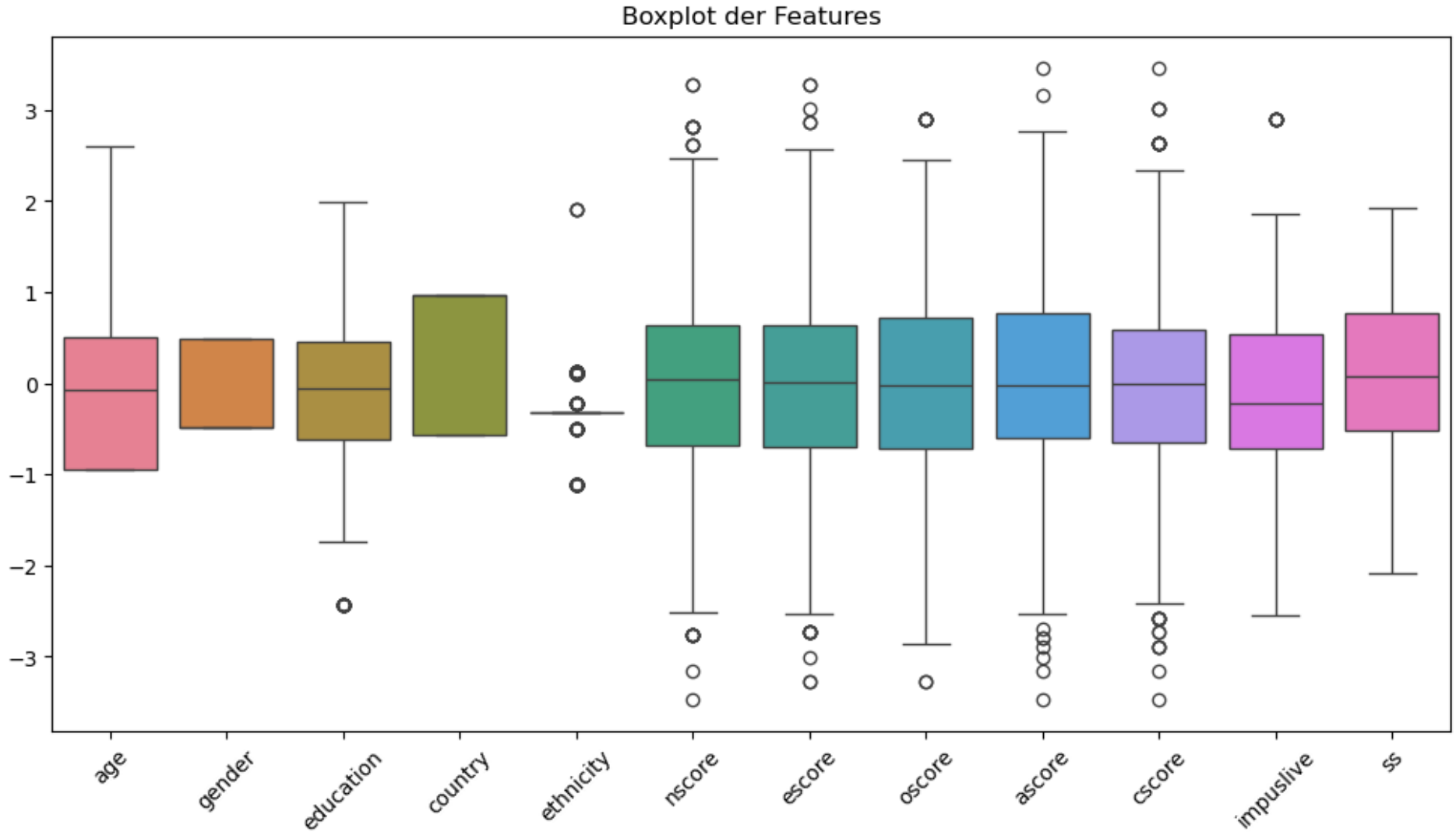
-> Führte in einem Drittprojekt bereits zu guten Ergebnissen

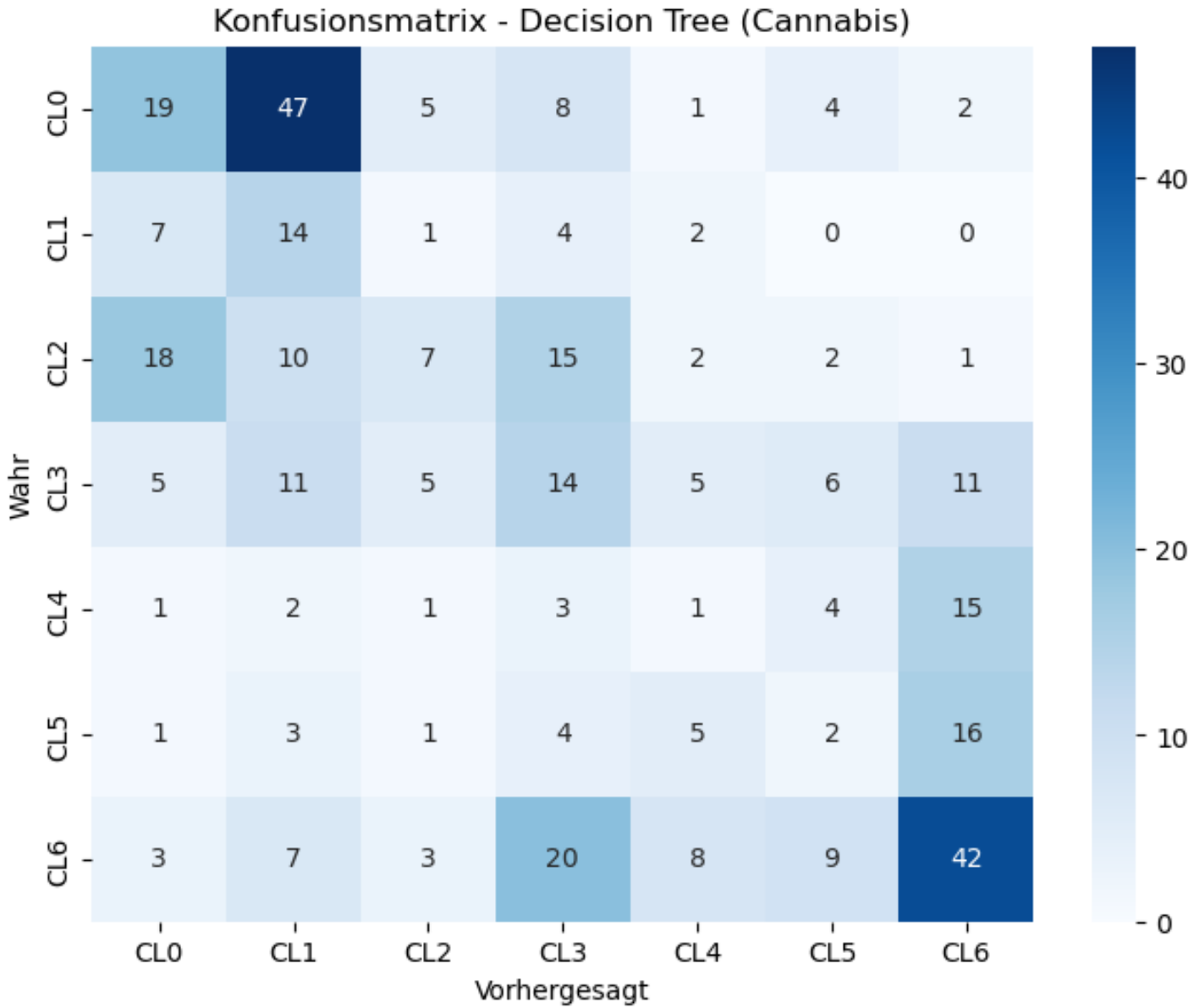
Sollte man die Multiklassifikation beibehalten wollen

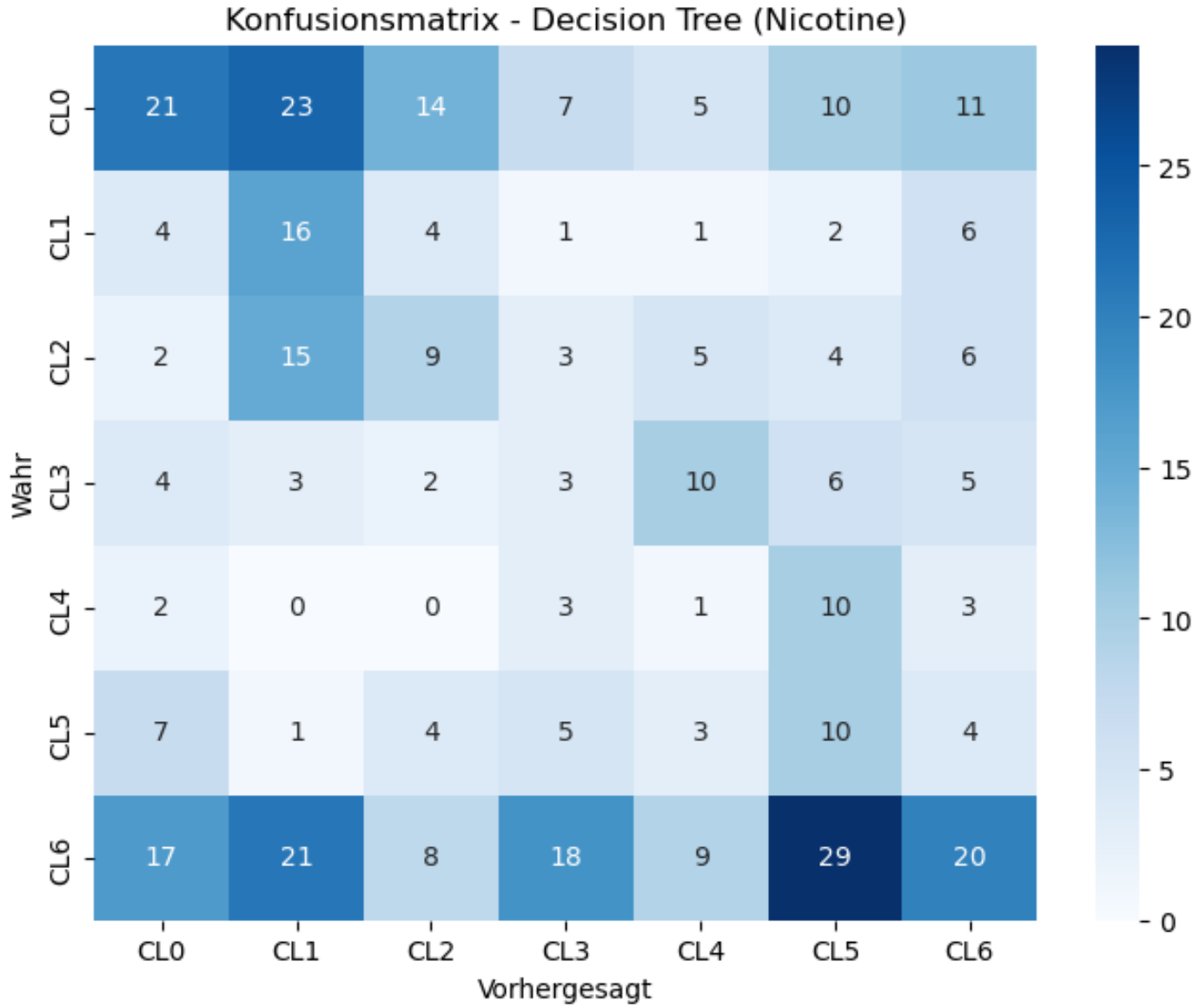
- Resampling des Datensatzes (SMOTE)
- Umfragen ausweiten -> Datenbasis vergrößern
- Angepassten F1-Score nutzen

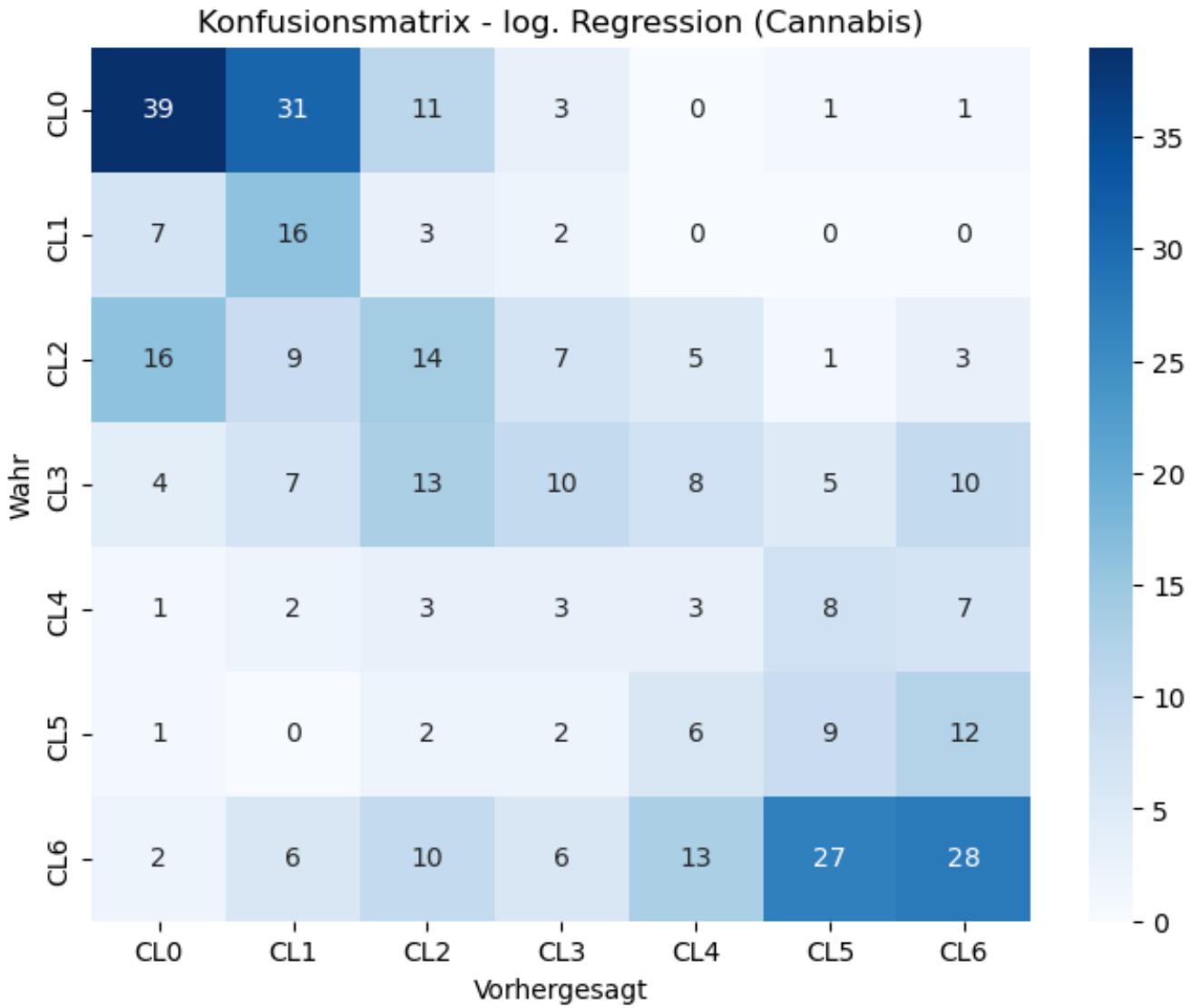
```
ACCURACY
Logisitc Regression Accuracy: 100.00%
Ridge Classifier Accuracy: 100.00%
Support Vector Machines Accuracy: 99.73%
Random Forest Classifier Accuracy: 100.00%
----------------------------------------
F1 SCORES
Logisitc Regression F1-Score: 1.0
Ridge Classifier F1-Score: 1.0
Support Vector Machines F1-Score: 0.99631
Random Forest Classifier F1-Score: 1.0
```

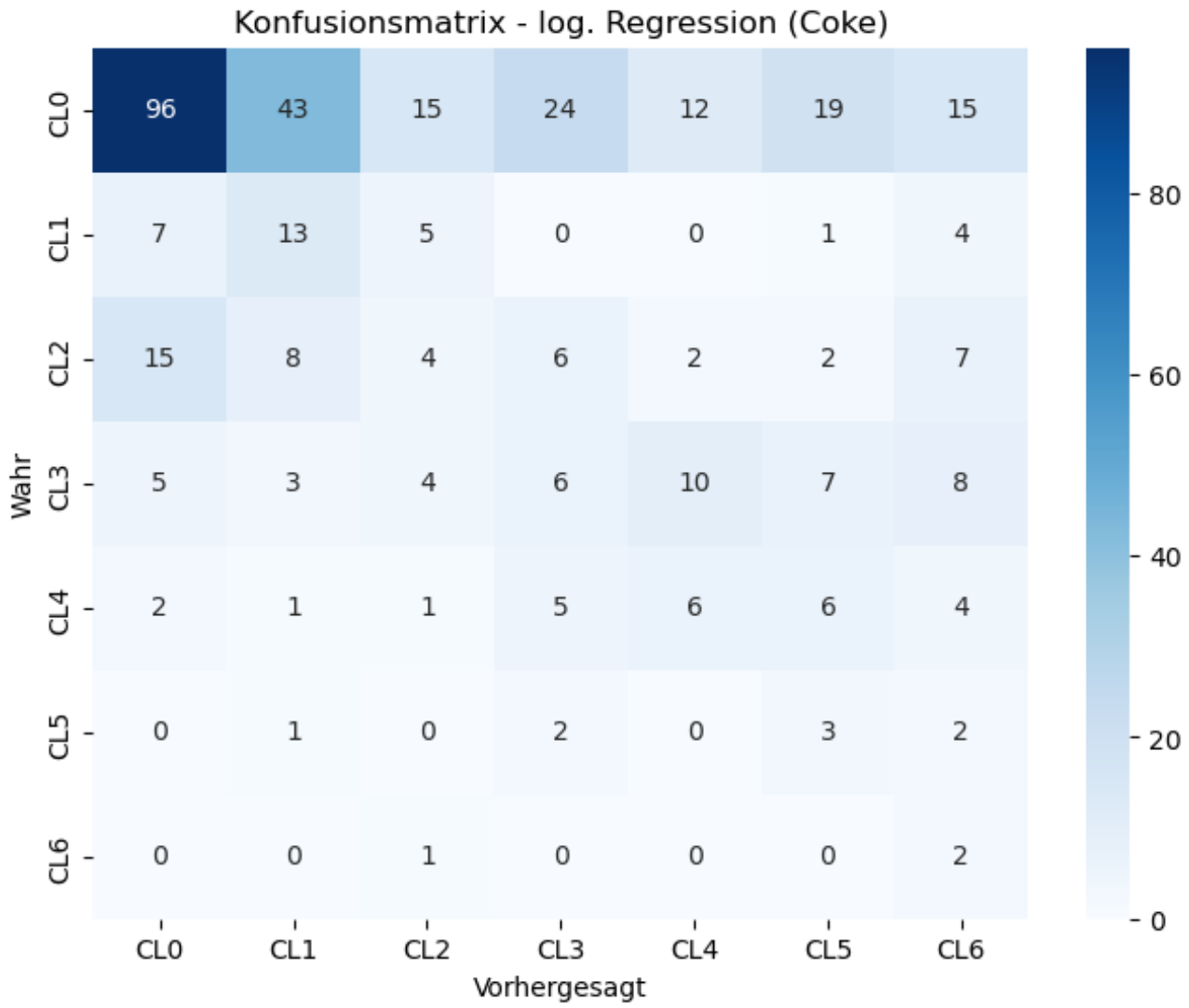
Fragen?

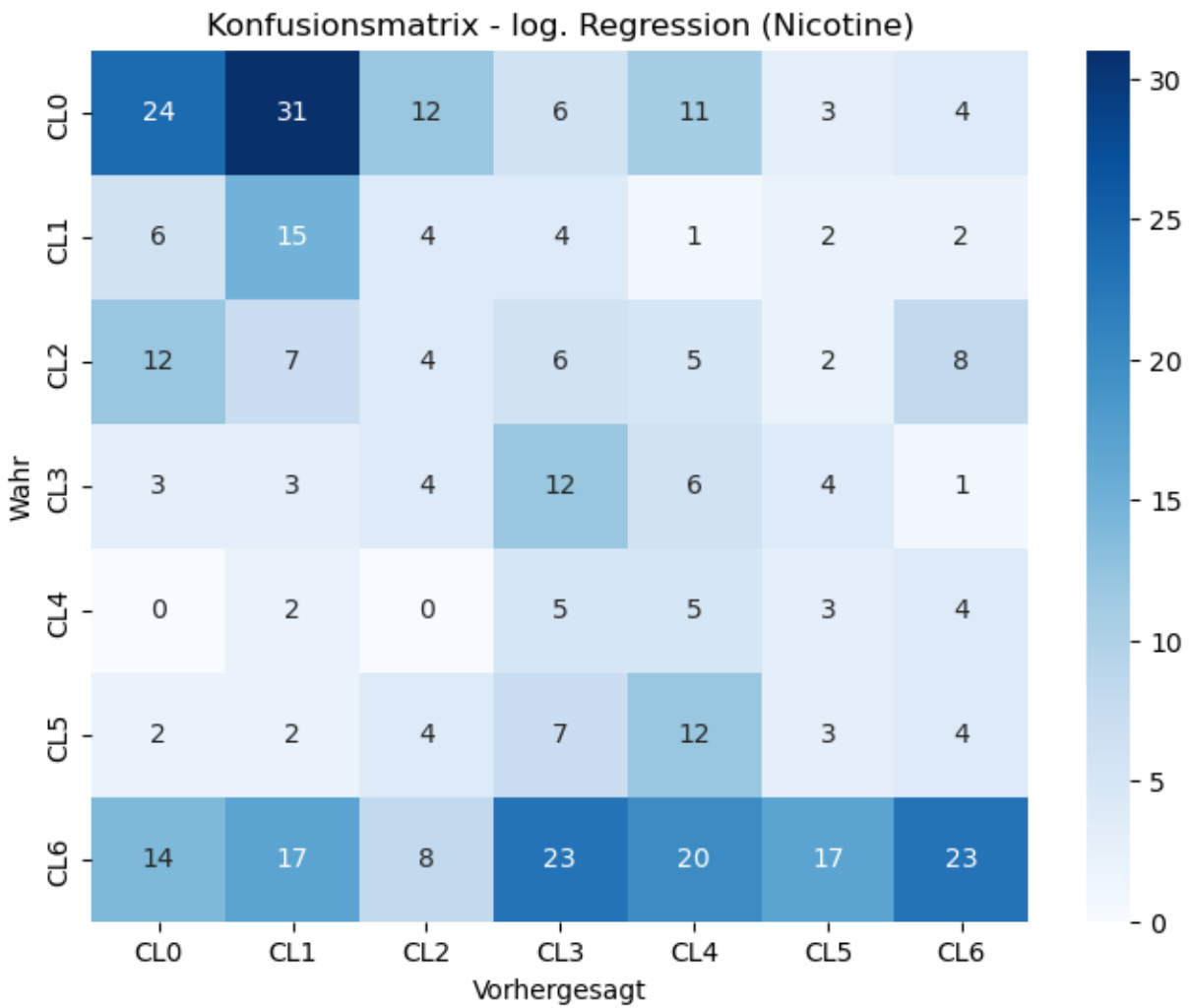

Boxplot der Features
3
2
1
0
−1
−2
−3
age
gender
education
country
ethnicity
nscore
escore
oscore
ascore
cscore
impuslive
ss


Konfusionsmatrix - Decision Tree (Cannabis)
Wahr
CL0
CL1
CL2
CL3
CL4
CL5
CL6
19 47 5 8 1 4 2
7 14 1 4 2 0 0
18 10 7 15 2 2 1
5 11 5 14 5 6 11
1 2 1 3 1 4 15
1 3 1 4 5 2 16
3 7 3 20 8 9 42
CL0
CL1
CL2
CL3
CL4
CL5
CL6
Vorhergesagt
40
30
20
10
0


Konfusionsmatrix - Decision Tree (Nicotine)
Wahr
Vorhergesagt
CL0
CL1
CL2
CL3
CL4
CL5
CL6
21 23 14 7 5 10 11
4 16 4 1 1 2 6
2 15 9 3 5 4 6
4 3 2 3 10 6 5
2 0 0 3 1 10 3
7 1 4 5 3 10 4
17 21 8 18 9 29 20
25
20
15
10
5
0


Konfusionsmatrix - log. Regression (Cannabis)
Wahr
CL0
CL1
CL2
CL3
CL4
CL5
CL6
39 31 11 3 0 1 1
7 16 3 2 0 0 0
16 9 14 7 5 1 3
4 7 13 10 8 5 10
1 2 3 3 3 8 7
1 0 2 2 6 9 12
2 6 10 6 13 27 28
CL0 CL1 CL2 CL3 CL4 CL5 CL6
Vorhergesagt
35
30
25
20
15
10
5
0


Konfusionsmatrix - log. Regression (Coke)
CL0
CL1
CL2
CL3
CL4
CL5
CL6
Wahr
96 43 15 24 12 19 15
7 13 5 0 0 1 4
15 8 4 6 2 2 7
5 3 4 6 10 7 8
2 1 1 5 6 6 4
0 1 0 2 0 3 2
0 0 1 0 0 0 2
CL0
CL1
CL2
CL3
CL4
CL5
CL6
Vorhergesagt
80
60
40
20
0


Konfusionsmatrix - log. Regression (Nicotine)
Wahr
CL0
CL1
CL2
CL3
CL4
CL5
CL6
24 31 12 6 11 3 4
6 15 4 4 1 2 2
12 7 4 6 5 2 8
3 3 4 12 6 4 1
0 2 0 5 5 3 4
2 2 4 7 12 3 4
14 17 8 23 20 17 23
CL0
CL1
CL2
CL3
CL4
CL5
CL6
Vorhergesagt
30
25
20
15
10
5
0