2025

Data Exploration Project

PREDICTION OF DRUG CONSUMPTION

SALIH KELMENDI, LUIS RASTETTER

WWI23B4 | 06.04.2025

Inhaltsverzeichnis

Abbildungsverzeichnis 2
1. Vorwort 3
2. Business Understanding 3
2.1 Auswahl des Datensatzes 3
2.2 Literaturrecherche 3
2.3 SMART-Ziel Formulierung 5
3. Data Understanding 6
3.1 Übersicht Datensatz 6
3.2 Relevante Merkmale 7
3.3 Lage- und Streuungsmaße 8
3.4 Datenvisualisierung 10
4. Data Preparation 13
4.1 Fehlwerte und Duplikate 13
4.2 Weitere Kriterien für Datenqualität 14
5. Modelling 15
5.1 Klärung der Art der Problemstellung 15
5.2 Erstellen der Modelle 15
6. Evaluation 20
6.1 Schwachstellen der Modelle 20
6.2 Auswahl des Modells 21
6.3 Merkmalsgewichtung 22
7. Ausblick 25

Abbildungsverzeichnis

Abbildung 1: SMART-Ziel 5
Abbildung 2: Übersichtstabelle Datensatz 7
Abbildung 3: Lage- und Streuungsmaße der Features 9
Abbildung 4: Violinplot der Features 9
Abbildung 5: Histogramme der Features 10
Abbildung 6: Verteilung des Konsums von Cannabis 11
Abbildung 7: Verteilung des Konsums von Nikotin 11
Abbildung 8: Verteilung des Konsums von Kokain 12
Abbildung 9: Bestätigung nicht vorhandener Null-Werte 13
Abbildung 10: Train-Test-Split 16
Abbildung 11: Optimierung Decision Tree mit Grid Search 17
Abbildung 12: Optimierter Decision Tree 18
Abbildung 13: Modellbewertung für Kokain 19
Abbildung 14: Modellbewertung für Nikotin 19
Abbildung 15: Modellbewertung für Cannabis 19
Abbildung 16: Konfusionsmatrix am Beispiel des Decision Trees und der Droge Kokain 21
Abbildung 17: Merkmalsgewichtung Decision Tree 22
Abbildung 18: Merkmalgewichtung Random Forest 23
Abbildung 19: Merkmalsgewichtung logistische Regression 23
Abbildung 20: Ergebnisse eines Drittprojekts 25

1. Vorwort

Dieses Portfolio wird im Rahmen der Vorlesung Data Exploration Project angefertigt. Beteiligte Personen an der Ausarbeitung sind Salih Kelmendi und Luis Rastetter.
Das Ziel dieser Prüfungsleistung ist das Trainieren und Vergleichen verschiedener Modelle, welche im Rahmen der Machine Learning Fundamentals im 3. Semester behandelt wurden, sodass die später aufgeführte Zielsetzung erreicht werden kann.
Der Vergleich jener Modelle erfolgt anhand der im folgenden Kapitel aufgestellten Ziel-Formulierung, welche einige Kriterien und den thematischen Zusammenhang liefert. Der gesamte Verlauf des Portfolios orientiert sich am Cross-Industry Standard Process for Data Mining.

2. Business Understanding

2.1 Auswahl des Datensatzes

Der ausgewählte Datensatz, mit welchem sich das Projekt beschäftigt, stammt aus dem UCI Machine Learning Repository und heißt „Drug Consumption (quantified)“ [1]. Er enthält Informationen über den Konsum verschiedener Drogen in Abhängigkeit von soziodemografischen und psychologischen Merkmalen.
Mit diesem soll die im Kapitel 2.3 aufgestellte SMART-Zielsetzung beantwortet werden.
Eine genauere Betrachtung und Auseinandersetzung mit dem Datensatz werden im weiteren Verlauf der Ausarbeitung folgen.

2.2 Literaturrecherche

Auf Basis des ausgewählten Datensatzes wurden bereits wissenschaftliche Paper und Studien erstellt. Hierbei fokussieren sich zwei Studien auf das Big-Five Modell aus der Persönlichkeitspsychologie[2], wie bestimmte Charaktereigenschaften mit dem Konsum von Drogen assoziiert sind. Die Studie „Big Five Personality Traits and Illicit Drug Use:

[1] https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

[2] https://linc.de/das-big-five-modell/

Specificity in Trait-Drug Associations”[3] beschreibt als Ergebnis, dass bestimmte Persönlichkeitsmerkmale mit dem Drogenkonsum zusammenhängen. So ist eine niedrige Gewissenhaftigkeit, beziehungsweise Verträglichkeit, als ein Indikator für einen hohen Drogenkonsum zu betrachten.

Eine zweite Studie „Big Five personality traits and alcohol, nicotine, cannabis, and gambling disorder comorbidity”[4] fand heraus, dass ein starker Alkoholkonsum mit Neurotizismus, also dass negative Emotionen wie Angst, Trauer, Nervosität und Reizbarkeit stärker erlebt werden, assoziiert, jedoch keine Assoziationen mit Extraversion, Offenheit für Neues und Verträglichkeit bestehen.

Zusätzlich soll auch eine generelle Studie für die Ursachen von Drogenkonsum hinzugezogen werden. Diese nennt als Hintergründe vor allem psychische, soziale und biologische Faktoren als auch die genetische Veranlagung eines jeden Menschen.[5]

Des Weiteren ist auf Kaggle auch ein bereits erstellter Code zu dem Thema zu finden, welcher Machine Learning Modelle hierfür trainiert. Dieser nutzt den F1-Score und die Accuracy als Metriken und die logistische Regression, den Ridge Classifier, den Random Forest Classifier und die Support Vector Machines als Modelle. Der Code verwendet dabei eine binäre Klassifikation für das Trainieren der Modelle.[6]

[3] https://pubmed.ncbi.nlm.nih.gov/34766786/
[4] https://pubmed.ncbi.nlm.nih.gov/31094546/
[5] https://dassuchtportal.de/drogensucht/ursachen/
[6] https://www.kaggle.com/code/obeykhadija/drug-consumption-prediction

2.3 SMART-Ziel Formulierung

Das Projekt und der weitere Verlauf der Ausarbeitung basieren auf folgender SMART-Zielsetzung, welche auch an die vorherigen genannten Paper und Studien anknüpft:
„Bis zum Abschluss des 4. Semesters soll ein Vergleich von unterschiedlichen Machine Learning Modellen erfolgen, um eine Prediction zu Wahrscheinlichkeiten und Intensität, inwiefern bestimmte Personen, basierend auf Umfragen zu Persönlichkeitsmerkmalen und -informationen, in der Zukunft dazu tendieren, welche Art von legalisierten als auch illegalen Drogen, zu konsumieren, treffen zu können. Der Erfolg der Modelle wird anhand von Trainings- und Testfehlern, F1-Score und Accuracy evaluiert, um das leistungsfähigste Modell für präzise Prognosen zu identifizieren.“

S **specific**	**M** **measurable**	**A** **achievable**	**R** **realistic**	**T** **time bound**
Entwicklung und Vergleich von ML-Modellen zur Prediction zum Konsum von Drogen	Messung der Modellqualitäten mit Accuracy, F1-Score, Test- und Trainingsfehler	Nutzbarer Datensatz für die Erstellung der Modelle in realistischem Projektzeitraum	Tangiert alltägliches Leben und kann auch für medizinische und Präventiv-maßnahmen relevant sein	Bis zum Ende des vierten Semesters

Abbildung 1: SMART-Ziel[7]

Zusätzlich zu der genannten Prediction soll auch die Inference im Projekt eine Rolle spielen. So soll ebenfalls eine Ursachenuntersuchung stattfinden, welche Merkmale besonders relevant für den Konsum von Drogen sind. Damit soll ein Vergleich zu den im Kapitel 2.2 genannten Ergebnisse der Studien erfolgen.

[7] Eigene Abbildung

3. Data Understanding

Nachdem im vorherigen Kapitel der Fokus auf der analytischen Zielsetzung des Projekts lag, soll nun ein erster Überblick über den vorhandenen Datenbestand erfolgen und daraus bereits erste Erkenntnisse über bestehende Zusammenhänge und das Qualitätsniveau des Drug Consumption Datensatzes gewonnen werden.

3.1 Übersicht Datensatz

Der Datensatz stammt aus einer Online-Umfrage, welche von Elaine Fehrmann in den Jahren 2011 und 2012 durchgeführt wurde. Insgesamt wurden 1885 valide Teilnehmenden aus verschiedenen Ländern, beispielsweise den USA, dem Vereinigten Königreich und Neuseeland, befragt. Darunter sind 943 Männer und 942 Frauen. Zudem wurde eine Ordinal/Nominal feature quantifiction angewendet.
Damit dies besser veranschaulicht werden kann, soll hierfür eine Übersichtstabelle für den Datensatz folgen. Die Tabelle enthält alle Merkmale, welche in Bezug auf die Bedeutung, Einheit und das Skalenniveau nochmals detaillierter aufgeführt werden.

Variable	Beschreibung	Skalenniveau	Typ
id	Eindeutige Identifikationsnummer	Nominal	Integer
age	Alter	Intervall (z-transformiert)	Feature
gender	Geschlecht	Intervall (z-transformiert)	Feature
education	Bildungsgrad	Intervall (z-transformiert)	Feature
country	Nationalität	Intervall (z-transformiert)	Feature
ethnicity	Ethnische Herkunft	Intervall (z-transformiert)	Feature
nscore	Neurotizismus	Intervall (z-transformiert)	Feature
escore	Extraversion	Intervall (z-transformiert)	Feature
oscore	Offenheit für Erfahrung	Intervall (z-transformiert)	Feature
ascore	Verträglichkeit	Intervall (z-transformiert)	Feature

cscore	Gewissenhaftigkeit	Intervall (z-transformiert)	Feature
impulsive	Impulsivität	Intervall (z-transformiert)	Feature
ss	Sensation-Suche	Intervall (z-transformiert)	Feature
alcohol	Konsum von Alkohol	Ordinal	Target
amphet	Konsum von Amphetamin	Ordinal	Target
amyl	Konsum von Amyl Nitrit	Ordinal	Target
benzos	Konsum von Benzos	Ordinal	Target
caff	Konsum von Koffein	Ordinal	Target
cannabis	Konsum von Cannabis	Ordinal	Target
choc	Konsumvon Schokolade	Ordinal	Target
coke	Konsum von Kokain	Ordinal	Target
crack	Konsum von Crack	Ordinal	Target
ecstasy	Konsum von Ecstacy	Ordinal	Target
heroin	Konsum von Heroin	Ordinal	Target
ketamine	Konsum von Ketamin	Ordinal	Target
legalh	Konsum von Legal Highs	Ordinal	Target
lsd	Konsum von LSD	Ordinal	Target
meth	Konsum	Ordinal	Target
mushrooms	Konsum von Pilzen	Ordinal	Target
nicotine	Konsum von Nikotin	Ordinal	Target
semer	Fiktive Droge	Ordinal	Target
vsa	Missbrauch flüchtiger Substanzen	Ordinal	Target

Abbildung 2: Übersichtstabelle Datensatz[8]

3.2 Relevante Merkmale

Wie in der vorherigen Übersichtstabelle des Datensatzes ersichtlich wird, besteht der Datensatz aus 12 Inputvariablen, welche die Hauptdimensionen der menschlichen Persönlichkeit aus dem Big Five-Modell der Persönlichkeitspsychologie als auch demografischen Merkmale der befragten Personen enthält. Diese sind für das Trainieren geeigneter Modelle alle relevant, weshalb vorerst keine der Features aus dem Datensatz entfernt werden.

[8] Eigene Abbildung

Zudem sind 19 Zielvariablen in Form von einzelnen Drogenarten, wie beispielsweise Alkohol oder LSD, enthalten. Da jedoch das Trainieren der jeweiligen Machine Learning Modelle auf die Vorhersage, dass eine Person in Zukunft eine dieser 19 Drogenarten konsumiert zu umfangreich wäre, werden für das weitere Vorgehen nur einzelne Zielvariablen in Betracht gezogen. Folglich werden die einzelnen Modelle hauptsächlich auf die Prediction des Konsums von Cannabis, Kokain und Nikotin trainiert.

3.3 Lage- und Streuungsmaße

Damit die einzelnen Modelle jedoch auch adäquat auf die ausgewählten Zielvariablen trainiert werden können, müssen zunächst die Lage- und Streuungsmaße der Inputvariablen im Datensatz überprüft werden. So muss überprüft werden, ob die Skalierung dieser angepasst werden muss und welche durchschnittliche Position die einzelnen Daten besitzen.

Die Betrachtung für den vorliegenden Datensatz ergibt, dass eine Anpassung der Skalierung jedoch nicht weiter notwendig ist und wird somit für den weiteren Verlauf des Trainierens der Modelle beibehalten. Ebenfalls ist eine gleichmäßige Verteilung der Features um einen zentralen Wert gegeben, sodass keine allzu robusten Modelle von Nöten sind, die eine besondere Variabilität berücksichtigen.

Die folgenden Abbildungen bestätigen diese Erkenntnisse. Die verschiedenen Scores aus dem Big-Five Modell sind weitestgehend normalverteilt und symmetrisch, ähnlich auch das Feature Sensation Seeking. Einzig und allein das Feature des Herkunftslandes ist um zwei Werte stark verteilt. Das begründet die hohe Menge an befragten Personen aus den USA (skalierter Wert von -0.57009) und dem Vereinigten Königreich (skalierter Wert von 0.96082).

Lage- und Streuungsmaße der Features:

	education	country	nscore	escore	oscore	\
count	1885.000000	1885.000000	1885.000000	1885.000000	1885.000000	
mean	-0.003806	0.355542	0.000047	-0.000163	-0.000534	
std	0.950078	0.700335	0.998106	0.997448	0.996229	
min	-2.435910	-0.570090	-3.464360	-3.273930	-3.273930	
25%	-0.611130	-0.570090	-0.678250	-0.695090	-0.717270	
50%	-0.059210	0.960820	0.042570	0.003320	-0.019280	
75%	0.454680	0.960820	0.629670	0.637790	0.723300	
max	1.984370	0.960820	3.273930	3.273930	2.901610	

	ascore	cscore	impuslive	ss
count	1885.000000	1885.000000	1885.000000	1885.000000
mean	-0.000245	-0.000386	0.007216	-0.003292
std	0.997440	0.997523	0.954435	0.963701
min	-3.464360	-3.464360	-2.555240	-2.078480
25%	-0.606330	-0.652530	-0.711260	-0.525930
50%	-0.017290	-0.006650	-0.217120	0.079870
75%	0.760960	0.584890	0.529750	0.765400
max	3.464360	3.464360	2.901610	1.921730

Abbildung 3: Lage- und Streuungsmaße der Features[9]

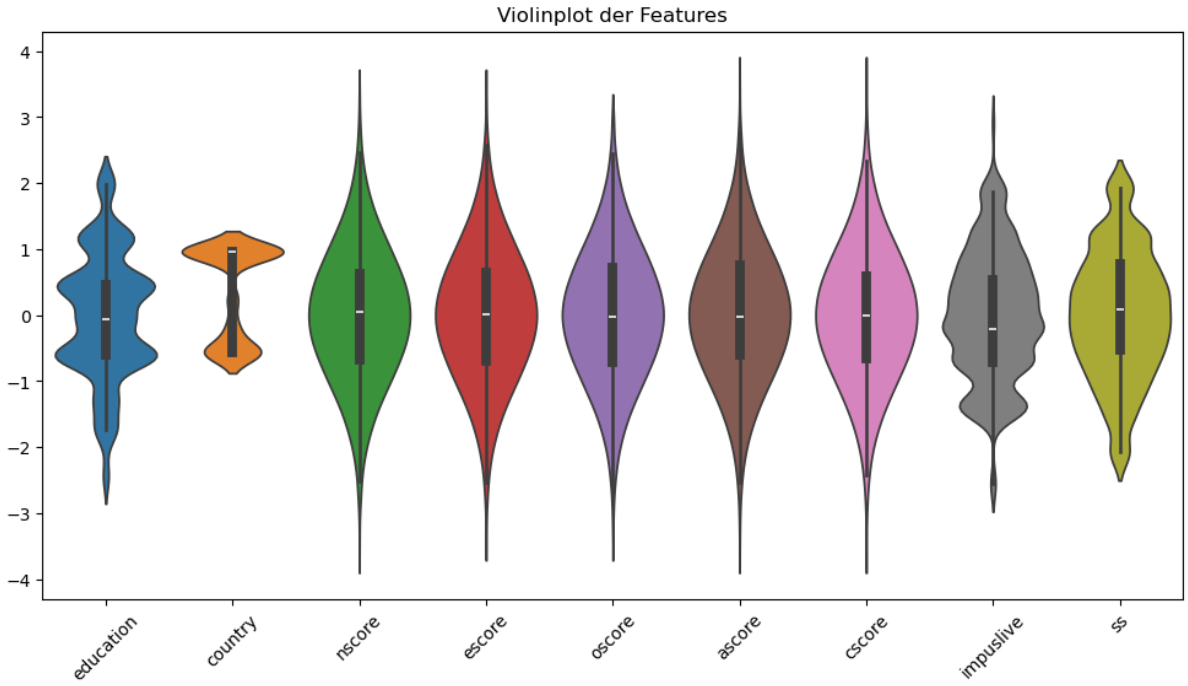


Abbildung 4: Violinplot der Features[10]

[9] Eigene Abbildung
[10] Eigene Abbildung

3.4 Datenvisualisierung

Um noch tiefgründigere Erkenntnisse über den Datensatz und dessen Inhalt zu erlangen, werden weitere Visualisierungen über die Verteilung der Features erstellt.

Als Erkenntnis hieraus lässt sich ebenso eine Gleichverteilung dieser, abgesehen von Spalten Country und Education, gewinnen (vgl. Abbildung 5).

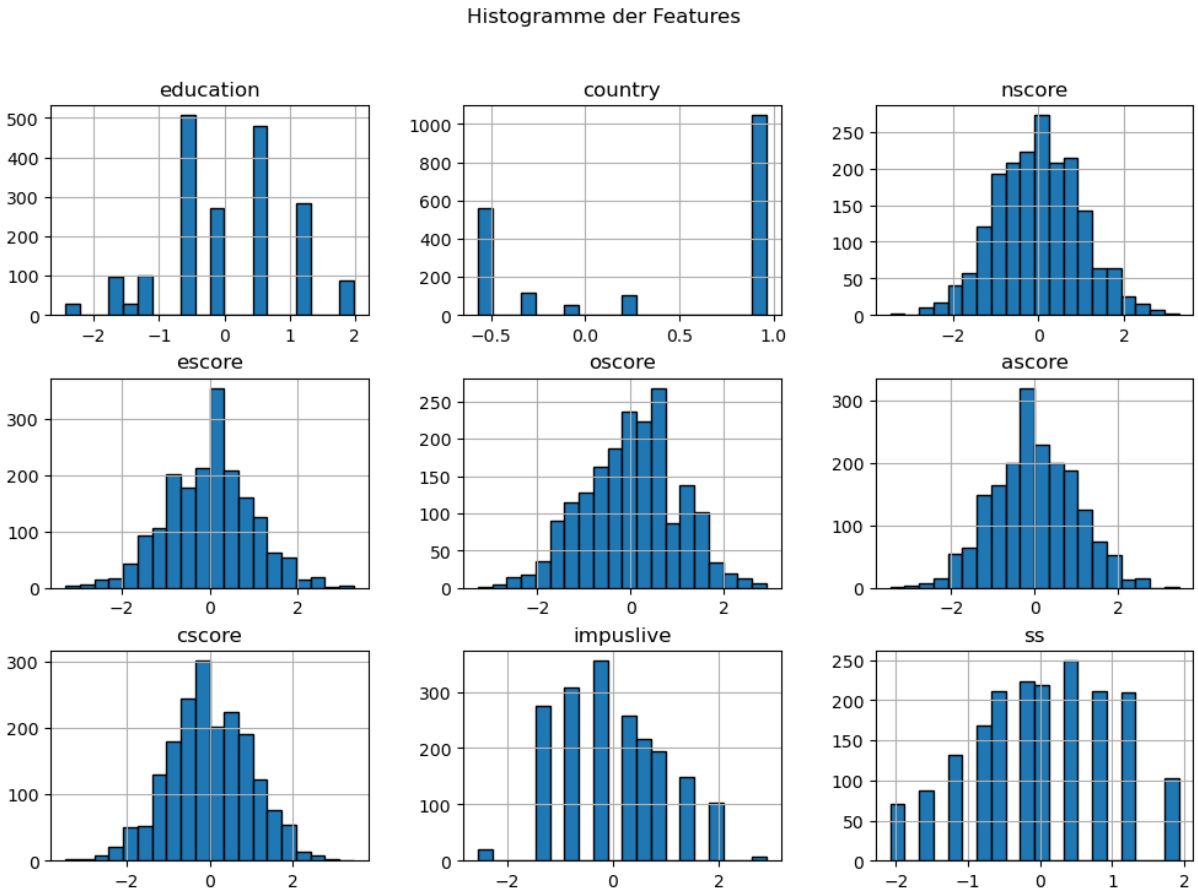


Abbildung 5: Histogramme der Features[11]

Außerdem sollen Grafiken für die Verteilung des Konsums von einzelnen Drogenarten erstellt werden. Hieraus lässt sich erkennen, dass der Konsum einer bestimmten Substanz in sieben verschiedene Kategorien unterteilt wird. Startend bei CL0 (noch nie konsumiert), über CL3 (im letzten Jahr konsumiert), bis hin zur Konsumklasse CL6, die eine Einnahme jener Droge am Tag vor der Befragung darstellt.

Zur Verdeutlichung werden hierfür die in Kapitel 3.2 ausgewählten Drogen Cannabis (vgl. Abbildung 6), Nikotin (vgl. Abbildung 7) und Kokain (vgl. Abbildung 8) verwendet.

[11] Eigene Abbildung

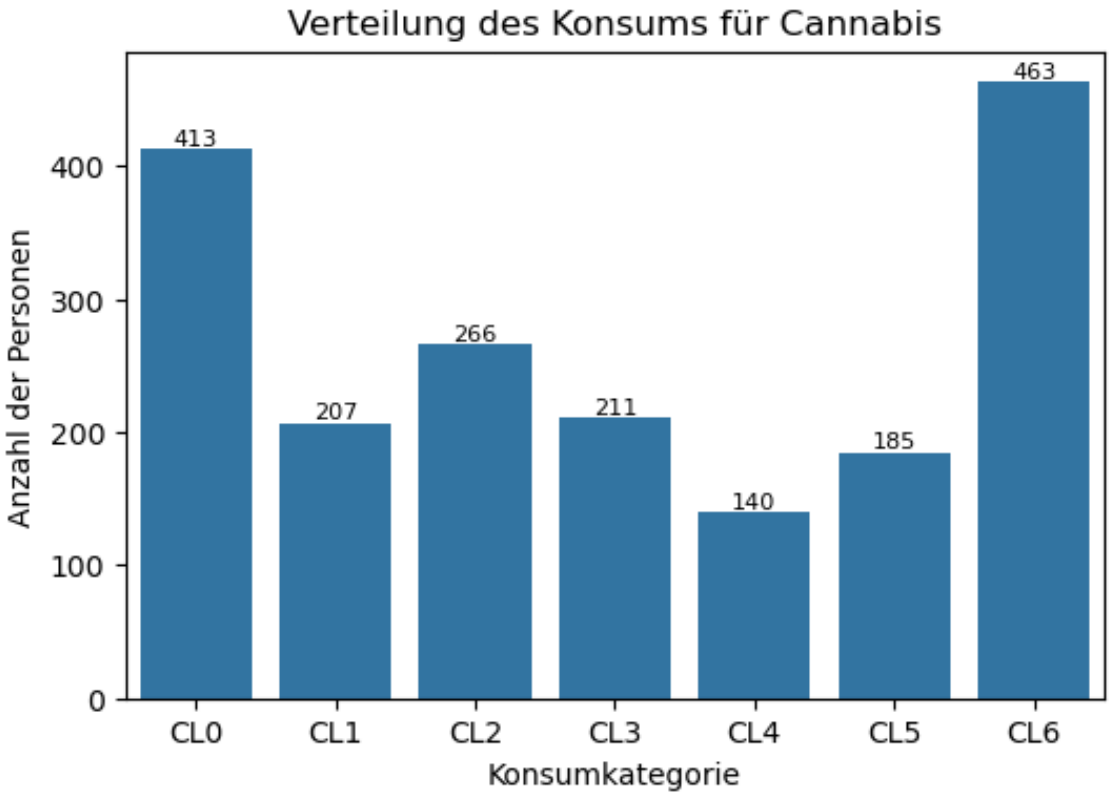


Abbildung 6: Verteilung des Konsums von Cannabis[12]

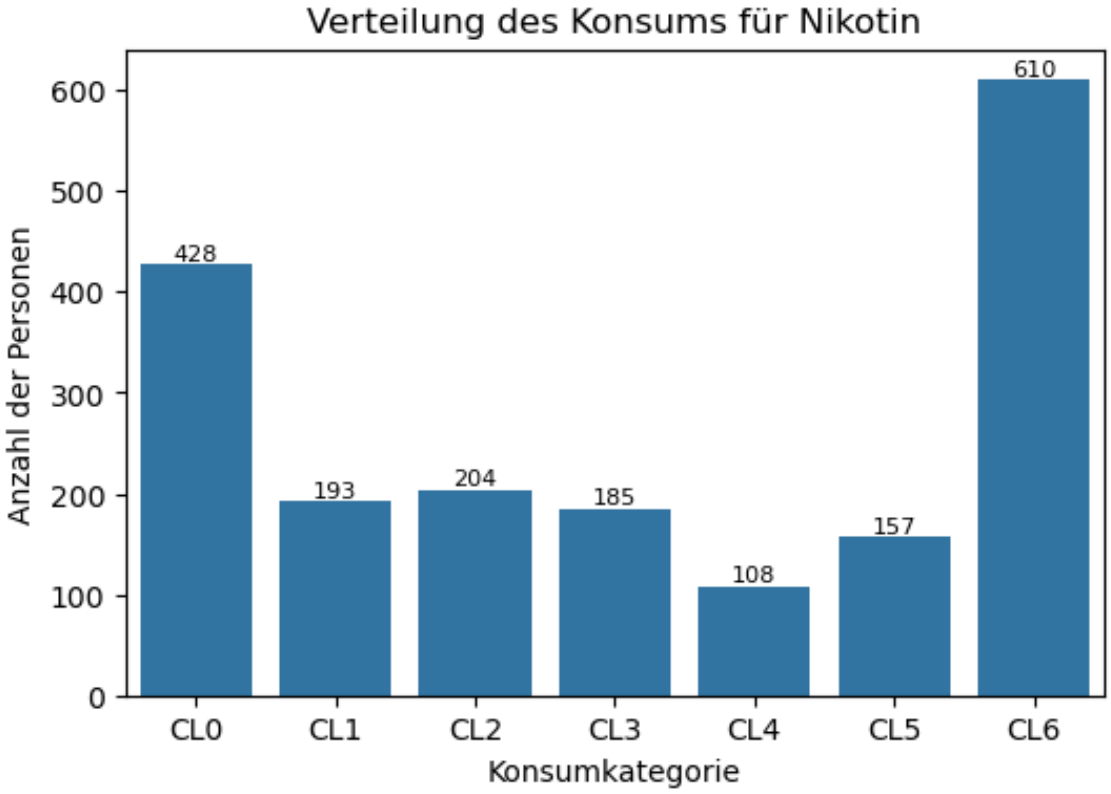


Abbildung 7: Verteilung des Konsums von Nikotin[13]

[12] Eigene Abbildung
[13] Eigene Abbildung

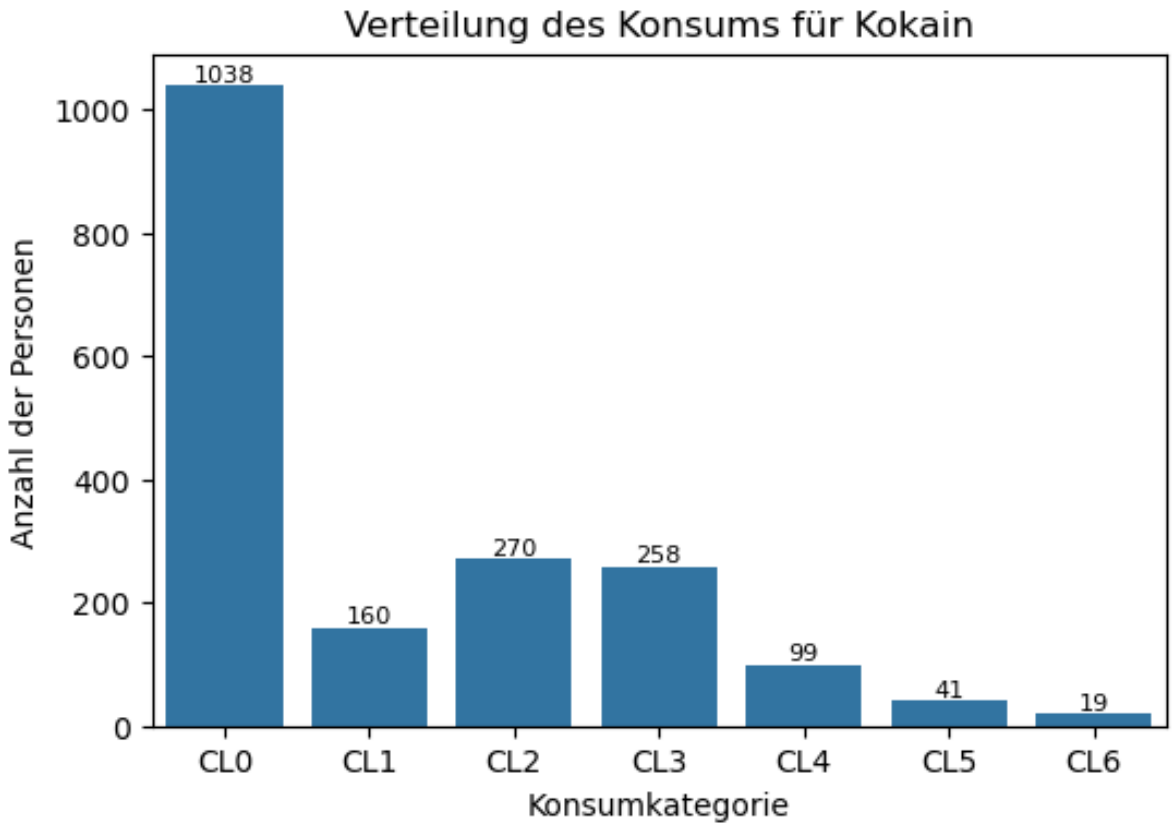


Abbildung 8: Verteilung des Konsums von Kokain[14]

Hierbei zeigen sich allerdings erste Probleme in Bezug auf die spätere Arbeit mit dem Datensatz. Für Drogen wie beispielsweise Kokain, welche im Vergleich zu Nikotin und Cannabis keine „Alltagsdrogen“ sind, enthält der Datensatz viel mehr Daten in der Konsumklasse CL0 als CL6, da bei diesen ein täglicher Konsum nahezu unmöglich ist. Dieser Aspekt wird im weiteren Verlauf des Modellings und der Evaluation ausführlicher behandelt

[14] Eigene Abbildung

4. Data Preparation

Nachdem im vorherigen Kapitel ein Überblick über den ausgewählten Datensatz geliefert wurde, soll dieser nun so vorbereitet werden, dass ein endgültiger Datensatz entsteht, welcher ein potenzielles Erstellen späterer Modelle ermöglicht.

4.1 Fehlwerte und Duplikate

Zuerst wird der Drug Consumption Datensatz auf mögliche Duplikate innerhalb der einzelnen Zeileneinträge überprüft, also ob eine Befragung einer bestimmten Person doppelt aufgeführt wird.

Ebenso muss untersucht werden, inwiefern die vorhandene Datenqualität für das weitere Vorgehen genügt. So dürfen keine Null- oder ähnliche Fehlwerte in den zur Benutzung vorgesehenen Datenzeilen vorhanden sein.

Nach Überprüfung dieser Kriterien ist festzustellen, dass keine Veränderungen am Datensatz notwendig werden und dieser nutzbar für die Ausarbeitung ist. Die vorhandenen numerischen Werte sind standardisiert mit einer Spanne zwischen einem Minimum und Maximum innerhalb eines vordefinierten Bereichs und es sind keine unlogischen als auch inkonsistenten Werte zu identifizieren.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1885 entries, 0 to 1884
Data columns (total 12 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        1885 non-null   float64
 1   gender     1885 non-null   float64
 2   education  1885 non-null   float64
 3   country    1885 non-null   float64
 4   ethnicity  1885 non-null   float64
 5   nscore     1885 non-null   float64
 6   escore     1885 non-null   float64
 7   oscore     1885 non-null   float64
 8   ascore     1885 non-null   float64
 9   cscore     1885 non-null   float64
 10  impuslive  1885 non-null   float64
 11  ss         1885 non-null   float64
dtypes: float64(12)
memory usage: 176.8 KB
```

Abbildung 9: Bestätigung nicht vorhandener Null-Werte[15]

[15] Eigene Abbildung

4.2 Weitere Kriterien für Datenqualität

Zusätzlich zur Bestätigung von nicht vorhandenen Fehlwerten und Duplikaten müssen noch weitere Kriterien, die eine ausreichende Datenqualität verifizieren, hinzugezogen werden. So ist eine normalisierte Verteilung der numerischen Daten vorauszusetzen, ohne dass zu viele Werte mit hohen positiven oder negativen Extremen auffallen.
Dieses Kriterium kann bereits mit der im Kapitel 3.3 aufgeführten Abbildung 2 bestätigt werden.

5. Modelling

Da nun nach dem Durchlaufen der vorherigen Prozessschritte des Data Understanding und der Data Preparation ausreichend Informationen über den Drug Consumption Datensatz vorhanden sind, soll in diesem Kapitel das Modelling im Vordergrund stehen.

5.1 Klärung der Art der Problemstellung

Bei der Betrachtung des vorliegenden Szenarios und der damit einhergehenden Problemstellung, ist zu erkennen, dass es sich hierbei um ein Klassifikationsproblem handelt. Der Output aus den zu trainierenden Modellen ist nicht numerisch, sondern qualitativ. Es soll anhand von diversen Features, welche in den vorherigen Kapiteln bereits aufgeführt wurden, versucht werden, vorherzusagen, ob eine Person in der Zukunft Drogen konsumiert.
Indiziert wird dies bereits, indem die befragten Personen bei der Umfrage angeben konnten, in welcher Klasse von CL0 bis CL6 sie sich einordnen würden. Hieraus resultiert, dass ein bestimmter Wert als Output auch einer dieser Klassen zugeordnet wird. Dementsprechend liegt sogar eine Multiklassifikation vor, da die trainierten Modelle nicht binär entscheiden können, ob eine Person Drogen konsumiert oder nicht. Die Modelle sollen mehrere Klassen voraussagen und Befragte in diese einordnen. Hierin soll sich das vorliegende Projekt vom anfangs aufgeführten Code in Kaggle unterscheiden, da in diesem die binäre Klassifikation zum Einsatz kommt.

5.2 Erstellen der Modelle

Zum Trainieren der einzelnen Modelle soll der Datensatz einen Train-Test-Split durchlaufen. Dieser sieht vor, dass die gesamte Datenmenge zu 80% in Trainingsdaten und zu 20% in Testdaten unterteilt wird. Dies entspricht 1508 Zeileneinträge für die Trainingsdaten und 377 für die Testdaten.

Abbildung 10: Train-Test-Split[16]

Zum Einsatz kommen während diesem Projekt die Modelltypen Decision Tree, KNN, SVM, logistische Regression als auch Random Forest. Allerdings wird der Fokus auf dem Decision Tree und der logistischen Regression liegen.

Der gesamte Prozess des Modellings besteht weitestgehend aus dem Wechseln zwischen Trainieren, Zwischenevaluierung und Optimieren der Modelle.

Da die trainierten Modelle auf die Drogen Nikotin, Cannabis und Kokain jedoch keine guten Ergebnisse liefern und die verwendeten Metriken F1-Score und Accuracy jeweils im Schnitt nur zwischen 0,2 und 0,4 liegen, dagegen die Test- und Trainingsfehler zu hoch sind, soll eine Hyperparameteroptimierung mittels Grid Search erfolgen, um mit Hilfe dessen eine Verbesserung der Modelle zu erwirken. Im Folgenden wird dies beispielhaft am Decision Tree aufgezeigt.

[16] Eigene Abbildung

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

# Parameter Grid für GridSearchCV
param_grid = {
    'max_depth': [3, 5, 10, 20, None],
    'max_features': [None, 'sqrt', 'log2'],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 5],
    'criterion': ['gini', 'entropy']
}

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))

    # DecisionTreeClassifier ohne Hyperparameter
    dt_model = DecisionTreeClassifier(random_state=42, class_weight=class_weight_dict)

    # GridSearchCV mit 5-Fold Cross-Validation
    grid_search = GridSearchCV(
        estimator=dt_model,
        param_grid=param_grid,
        cv=5,  # Cross-Validation mit 5 Folds
        n_jobs=-1,  # Nutzen aller verfügbaren CPU-Kerne
        verbose=1,  # Protokollierung des Fortschritts
        scoring='accuracy'  # Füge ein Scoring-Metrik hinzu (hier 'accuracy')
    )

    # Führe GridSearchCV aus
    grid_search.fit(X_train, y_train[drug])

    # Bestes Modell speichern
    best_dt_model = grid_search.best_estimator_
    dt_models[drug] = best_dt_model

    # Vorhersagen mit dem besten Modell
    y_pred_dt[drug] = best_dt_model.predict(X_test)

    # Ausgabe der besten Parameter
    print(f"Beste Hyperparameter für {drug}: {grid_search.best_params_}")
```

Abbildung 11: Optimierung Decision Tree mit Grid Search[17]

Aus dieser Optimierung ergibt sich der Codeabschnitt auf der nachfolgenden Seite.

[17] Eigene Abbildung

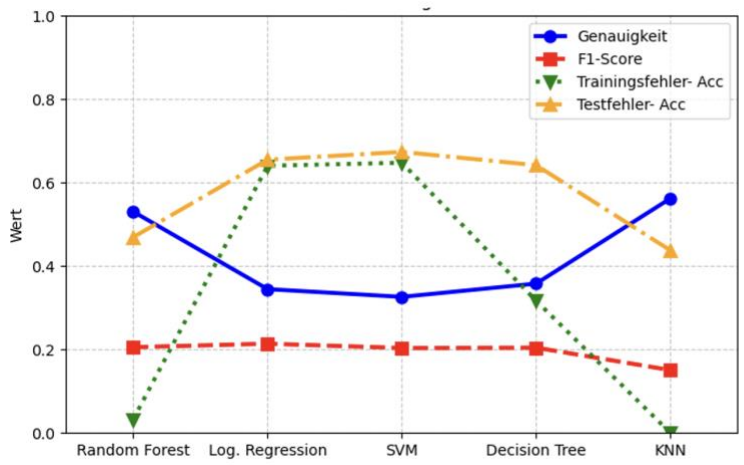
```python
dt_models = {}
y_pred_dt = {}

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,
        max_features='sqrt' if drug == 'nicotine' else None,
        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)
```

Abbildung 12: Optimierter Decision Tree[18]

Die erwartete Verbesserung der Modelle nach der Hyperparameteroptimierung ist jedoch nicht im erhofften Umfang zu erkennen. Die folgenden Abbildungen sollen die jeweiligen Modellbewertungen für die ausgewählten Drogenarten veranschaulichen. Ersichtlich wird, dass vor allem bei jenen Drogenarten wie Kokain, Konsumklassen, die nur wenige Daten enthalten, nur wenig gut bis überhaupt nicht vorausgesagt werden und dementsprechend auch nur eine geringe Accuracy und ein geringer F1-Score damit einhergehen (vgl. Abbildung 13). Dies ist sowohl bei der logistischen Regression als auch dem Decision Tree der Fall.

Auch im Falle der beiden weiteren Drogenarten fallen die Metriken nicht zufriedenstellend aus, da es für die Modelle schlichtweg schwer ist, vor allem bei einer nur geringen Datenmenge für die einzelnen Konsumklassen, die Daten richtig in die sieben verschiedenen Konsumklassen zuzuordnen (vgl. Abbildung 14 und 15).

[18] Eigene Abbildung



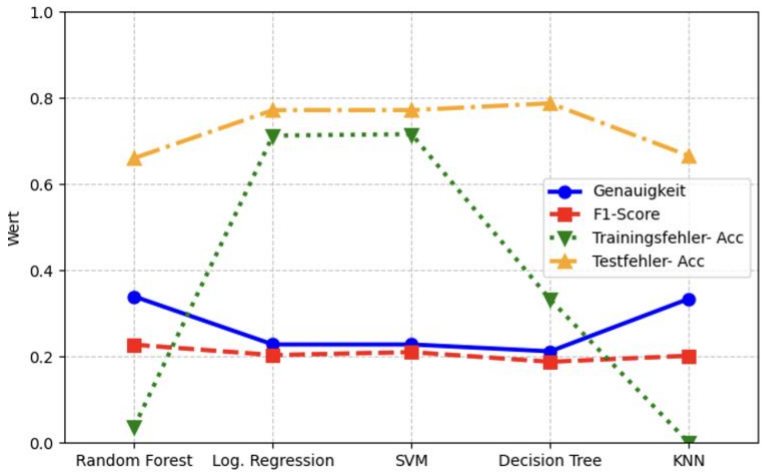
Class	Precision	Recall	F1-Score	Support
CL0	0,71	0,44	0,54	224
CL1	0,28	0,43	0,34	30
CL2	0,18	0,25	0,21	44
CL3	0,17	0,21	0,19	43
CL4	0,07	0,12	0,09	25
CL5	0,04	0,12	0,06	8
CL6	0	0	0	3
accuracy		0,36		377
macro avg	0,21	0,23	0,2	377
weighted avg	0,49	0,36	0,4	377

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

Abbildung 13: Modellbewertung für Kokain[19]



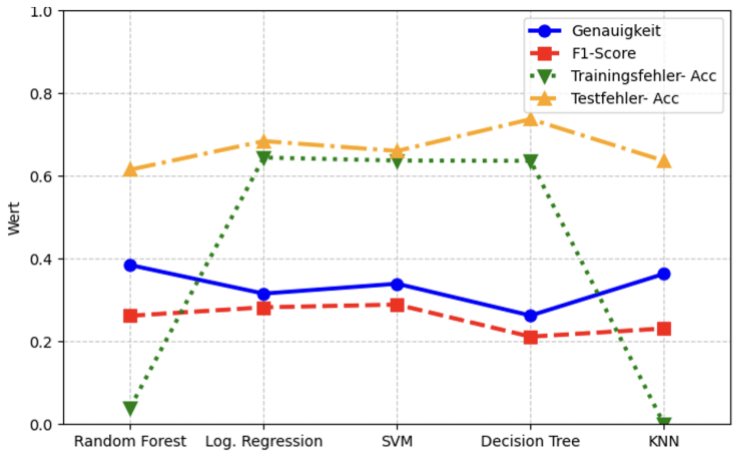
Class	Precision	Recall	F1-Score	Support
CL0	0,37	0,23	0,28	91
CL1	0,2	0,47	0,28	34
CL2	0,22	0,2	0,21	44
CL3	0,07	0,09	0,08	33
CL4	0,03	0,05	0,04	19
CL5	0,14	0,29	0,19	34
CL6	0,36	0,16	0,23	122
accuracy		0,21		377
macro avg	0,2	0,22	0,19	377
weighted avg	0,27	0,21	0,22	377

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

Abbildung 14: Modellbewertung für Nikotin[20]



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

Abbildung 15: Modellbewertung für Cannabis[21]

[19] Eigene Abbildung
[20] Eigene Abbildung
[21] Eigene Abbildung

6. Evaluation

Nachdem im vorhergehenden Kapitel die beschriebenen Modelle trainiert und modelliert wurden, sollen diese nun abschließend verglichen und bewertet werden. Zudem soll aufgezeigt werden, welche Schwachstellen die verwendeten Modelle besitzen.

6.1 Schwachstellen der Modelle

Wie bereits im Kapitel 5 des Modellings erwähnt wurde, verzeichnen die erstellten Modelle keine guten Werte im Hinblick auf die gewählten Metriken.
Dies hängt mit der vorliegenden Class Imbalance zusammen. In den Testdaten sind viel mehr Daten, die der Konsumklasse CL0 zugeordnet werden als jene, die in die Klasse CL6 klassifiziert werden, zu finden. Deshalb werden die Modelle alle stark in Richtung der Klasse CL0 optimiert, wodurch die Schwierigkeit entsteht, dass seltene Klassen wie Cl6 korrekt klassifiziert werden können. Anhand der Abbildungen acht, 13 und 16 lässt sich dies sehr gut erkennen. Für Kokain sind in den Klassen CL4, CL5 und CL6 nur sehr wenige Daten vorhanden, weshalb diese von keinem Modell adäquat klassifiziert werden können.
Ebenso ist die Multiklassifikation an sich ein Schwachpunkt der einzelnen Modelle, da diese nicht binär klassifizieren, sondern zwischen vielen Klassen unterscheiden müssen. So können sich Konsumklassen wie CL2, CL3 und CL4 in den Merkmalen stark überschneiden, was zu Fehlklassifikationen führt.
Des Weiteren ist der Drogenkonsum an sich auch sehr komplex und schwer vorhersagbar, da auch abgesehen von den im Datensatz enthaltenen Features diverse weitere Faktoren in das Konsumieren von Rauschmitteln mit einfließen.
All diese Aspekte sind ursächlich dafür, dass die folgende Konfusionsmatrix für den Decision Tree am Beispiel Kokain wenig richtig erscheinen mag.

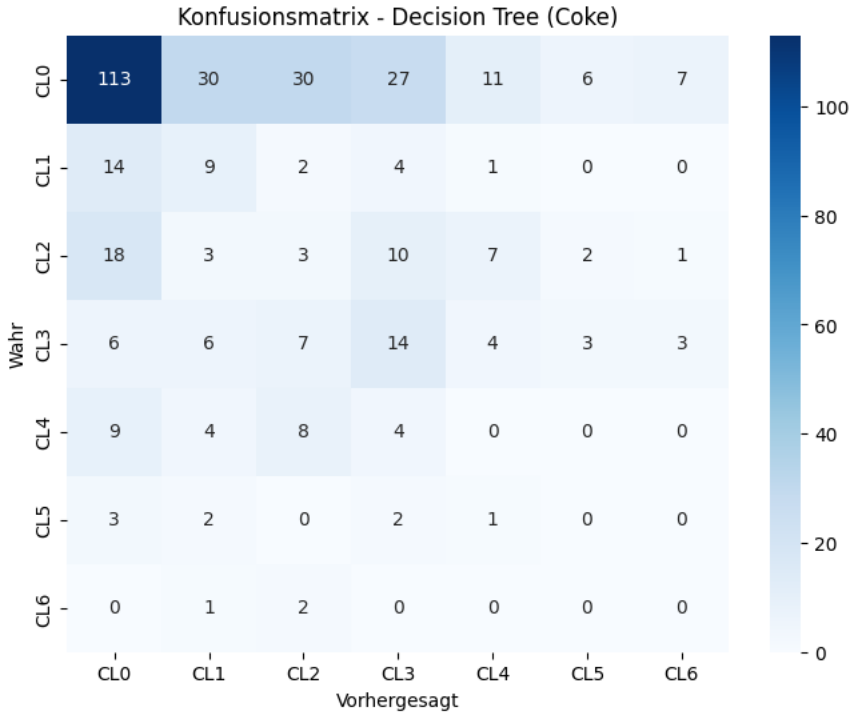


Abbildung 16: Konfusionsmatrix am Beispiel des Decision Trees und der Droge Kokain[22]

6.2 Auswahl des Modells

Grundsätzlich könnte man sich abschließend für ein Modell entscheiden, das für die Klassifikation des Drogenkonsums verwendet wird. In diesem Fall wäre die logistische Regression dem Decision Tree vorzuziehen, da diese in der Auswertung einen geringeren Testfehler und eine bessere Generalisierungsleistung zeigt. Auch in Bezug auf die Accuracy und den F1-Score zeigt diese bessere Werte und ist auch robuster den vorliegenden unausgeglichenen Klassen.
Weiter ist die logistische Regression weniger anfällig für Overfitting, da der Test- und Trainingsfehler in etwa gleich hoch sind und auch leichter interpretier- wie auch trainierbar als ein Decision Tree ist.

Auf der einen Seite gibt es ebenjene Argumente, um die Entscheidung für ein Modell zu treffen. Auf der anderen Seite erscheint es unter den Umständen der im Kapitel 6.1 aufgeführten Aspekte allerdings auch wenig sinnvoll ein Modell zu bevorzugen, da die Vorteile der logistischen Regression minimal sind und die Vergleichsmetriken auf einem

[22] Eigene Abbildung

ähnlich schlechten Niveau liegen. Um bessere Ergebnisse für das Vorhersagen der Klassen zu erzielen, müsste zunächst die Datenbasis angepasst und verbessert werden, da ansonsten diverse weitere Modelle auf die bestehenden Daten trainiert werden könnten, jedoch durch die Class Imbalance und die Multiklassifikation keine genaueren Ergebnisse erzielt werden würden.

6.3 Merkmalsgewichtung

Auch wenn die besagten Modelle keine geeigneten Ergebnisse erzielen, soll dennoch eine Merkmalsgewichtung für die einzelnen Features erfolgen, um die Ergebnisse aus den zu Beginn aufgeführten Studien vergleichen und die Inference für den Drogenkonsum begründen zu können. Bei einer neuen Datenbasis müsste diese zwar nochmals neu bewertet werden, dennoch besitzt diese trotz schlechter Modellergebnisse auch eine gewisse Aussagekraft.

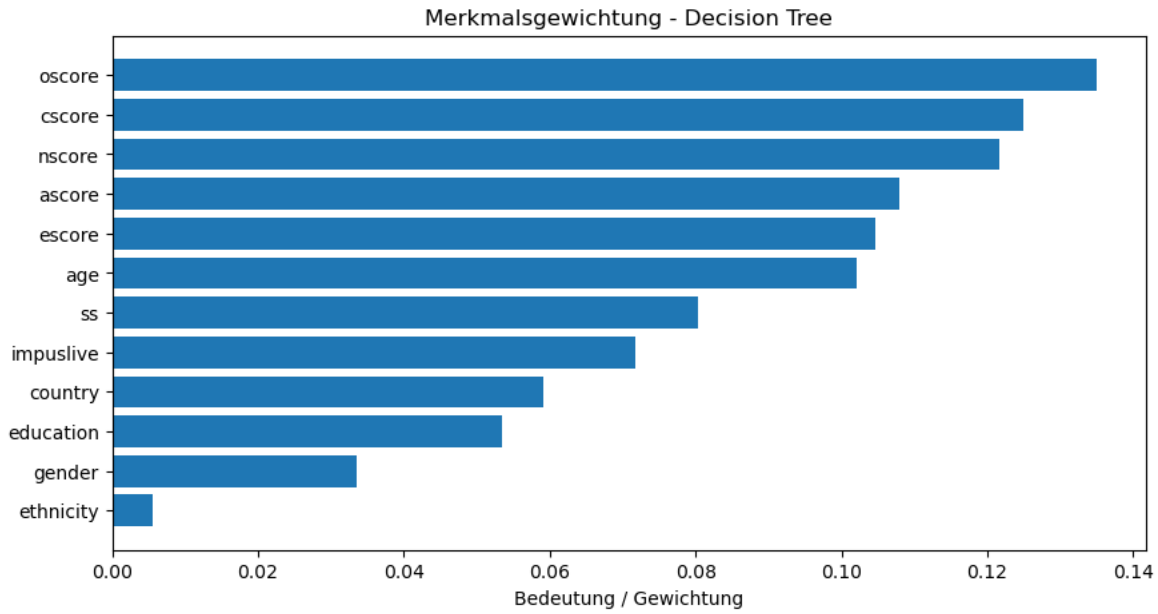


Abbildung 17: Merkmalsgewichtung Decision Tree[23]

[23] Eigene Abbildung

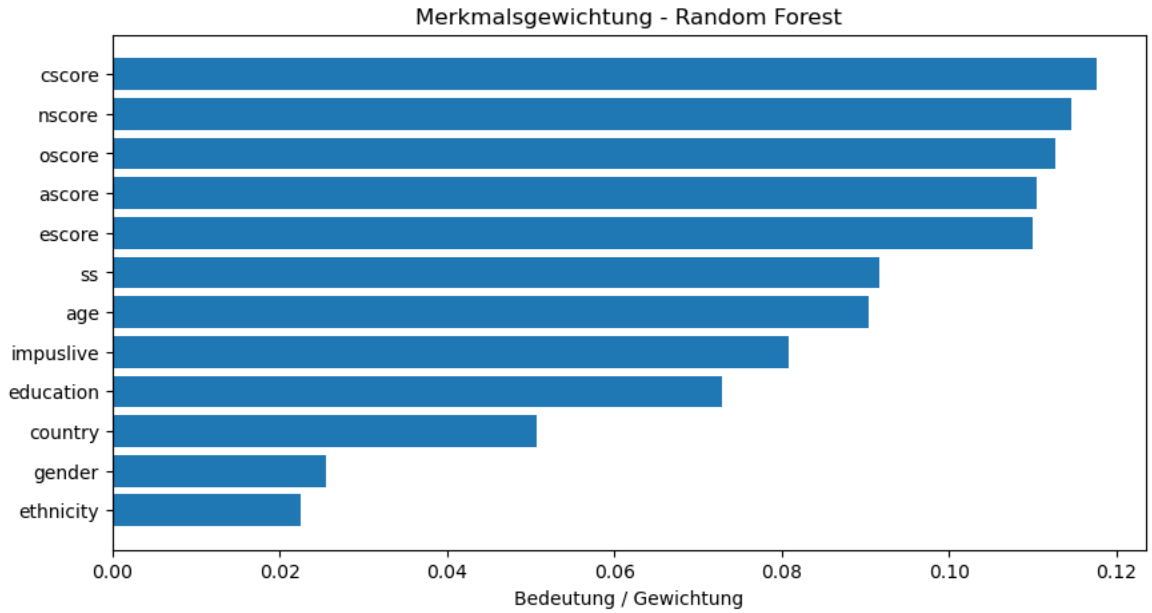


Abbildung 18: Merkmalgewichtung Random Forest[24]

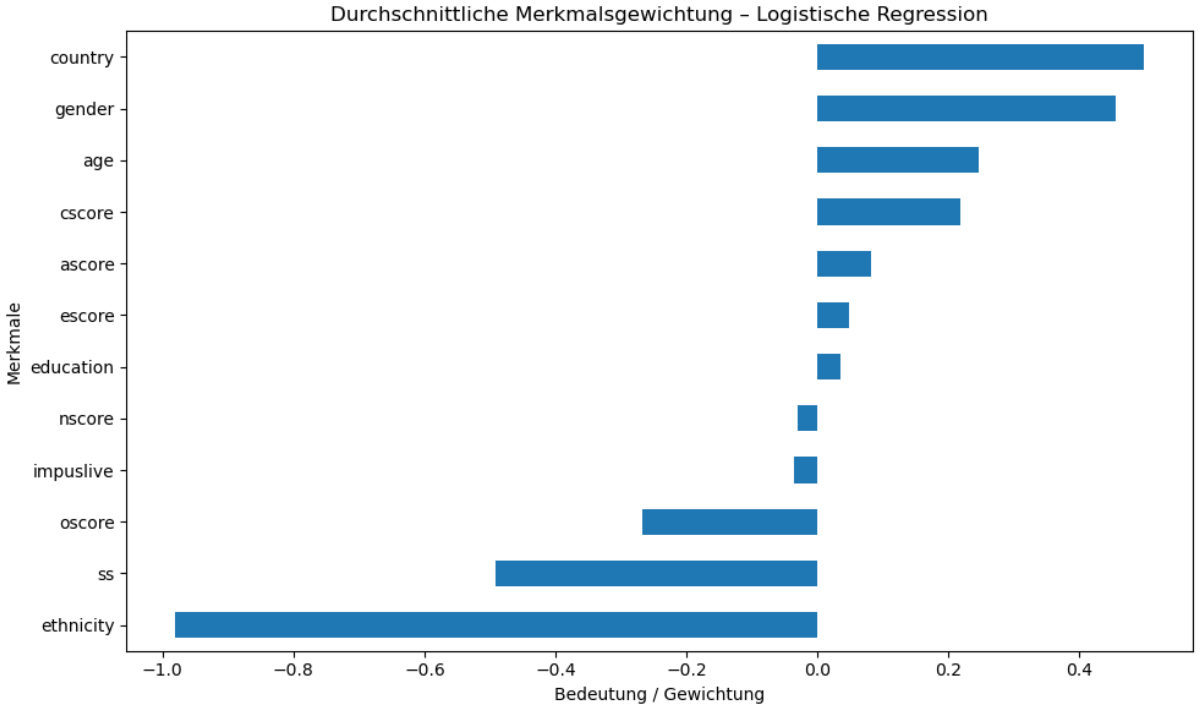


Abbildung 19: Merkmalsgewichtung logistische Regression[25]

Es zeigt sich, dass in jedem Modell die verschiedenen Scores aus dem Big Five-Modell, abgesehen von o- und nscore bei der logistischen Regression, eine einflussreiche Rolle spielen. Dadurch wird die Literatur bestätigt, indem gezeigt wird, dass psychologische

[24] Eigene Abbildung
[25] Eigene Abbildung

Faktoren einer der Hauptgründe sind, wieso Menschen in verschiedenen Ländern Drogen konsumieren.

Ebenfalls besitzt das Feature country bei der logistischen Regression eine hohe Gewichtung. Dieses müsste jedoch tiefgründiger auf die einzelnen Drogenbestimmungen in den jeweiligen Ländern analysiert werden, da beispielsweise die USA eine andere Gesetzesbestimmung zu legalen und illegalen Rauschmitteln vorweist als das Vereinigte Königreich und so der Drogenkonsum der jeweiligen Einwohner auf einer anderen Ebene beeinflusst wird.

7. Ausblick

Für die Zukunft können verschiedene Änderungen vorgenommen werden, um das Ergebnis der trainierten Modelle zu verbessern. Diese sind im Rahmen dieses Projektes allerdings nicht umzusetzen, da sie zu umfangreich wären.

So bestünde die Möglichkeit, die vorliegende Multiklassifikation zu einer binären Klassifikation zu ändern. Der zu Beginn im Kapitel 2.2 aufgeführte Code von Kaggle erzielte damit bereits sehr gute Ergebnisse (vgl. Abbildung 20).

Andernfalls, insofern die Multiklassifikation beibehalten werden soll, ist es möglich ein Resampling des Datensatzes, beispielsweise mit der Synthetic Minority Oversampling Technique (SMOTE), durchzuführen. Zusätzlich müsste die Menge an befragten Personen generell ausgeweitet werden, um eine geeignete Datenbasis zu erhalten, welche für die jeweiligen Konsumklassen ausreichend Daten aufweist.

Damit könnte abschließend erreicht werden, dass die Modelle für spätere Präventionsmaßnahmen gegen den Konsum von Drogen eingesetzt werden, da ersichtlich wird welche Zielklassen in der Bevölkerung angesprochen werden müssen und somit eine tatsächliche Anwendung in der Realität erfolgt.

```
                    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
```

Abbildung 20: Ergebnisse eines Drittprojekts[26]

[26] https://www.kaggle.com/code/obeykhadija/drug-consumption-prediction