Data Exploration Project

wwi23b4 | 06.04.2025

prediction of drug consumption

salih kelmendi, Luis Rastetter

2025

# Inhaltsverzeichnis

[1. Vorwort 2](#_Toc191463316)

[2. Business Understanding 2](#_Toc191463317)

[2.1 Auswahl des Datensatzes 2](#_Toc191463318)

[2.2 SMART-Ziel Formulierung 2](#_Toc191463319)

[2.3 Literaturrecherche 2](#_Toc191463320)

[3. Data Understanding 3](#_Toc191463321)

[3.1 Übersichtstabelle Datensatz 3](#_Toc191463322)

[3.2 Relevante Merkmale 4](#_Toc191463323)

[3.3 Lage- und Streuungsmaße 4](#_Toc191463324)

[3.4 Datenvisualisierung 4](#_Toc191463325)

[4. Data Preparation 5](#_Toc191463326)

[4.1 Fehlwerte 5](#_Toc191463327)

[4.2 Duplikate 5](#_Toc191463328)

[4.3 Weitere Kriterien für Datenqualität 5](#_Toc191463329)

[4.4 Skalierung und Standardisierung 5](#_Toc191463330)

<https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified>

# 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.

Das 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.

# Business Understanding

## 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]](#footnote-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 wird im weiteren Verlauf der Ausarbeitung folgen.

## Literaturrecherche

Auf Basis des ausgewählten Datensatzes wurden bereits wissenschaftliche Paper erstellt. Beispielsweise jenes mit dem Titel „The Five Factor Model of personality and evaluation of drug consumption risk“.[[2]](#footnote-2) Dieses beschreibt das Problem der Bewertung des individuellen Risikos für Drogenkonsum und -missbrauch.

Hierbei werden mannigfaltige Ansätze zur Lösung des Problems geliefert, an welchen auch die für die Ausarbeitung gewählte Zielsetzung und damit das Projekt im Gesamten anknüpfen sollen und sich daran orientieren.

## SMART-Ziel Formulierung

Das Projekt und der weitere Verlauf der Ausarbeitung basiert auf folgender SMART-Zielsetzung, welche an das vorherige genannte Paper 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.“*

# 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.

## Übersichtstabelle Datensatz

Hierfür soll zunächst eine Übersichtstabelle für den 1885 Zeilen großen Datensatz, wovon jede Zeile eine befragte Person zu ihrem Drogenkonsum in der Vergangenheit darstellt, angefertigt werden. 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 | Ordinal | Feature |
| gender | Geschlecht | Nominal | Feature |
| education | Bildungsgrad |  | Feature |
| country | Nationalität |  | Feature |
| ethnicity | Ethnische Herkunft |  | Feature |
| nscore | Neurotizismus |  | Feature |
| escore | Extraversion |  | Feature |
| oscore | Offenheit für Erfahrung |  | Feature |
| ascore | Verträglichkeit |  | Feature |
| cscore | Gewissenhaftigkeit |  | Feature |
| impulsive |  |  | Feature |
| ss |  |  | Feature |
| alcohol | Konsum von Alkohol |  | Target |
| amphet | Konsum von Amphetamin |  | Target |
| amyl | Konsum von Amyl Nitrit |  | Target |
| benzos | Konsum von Benzos |  | Target |
| caff | Konsum von Koffein |  | Target |
| cannabis | Konsum von Cannabis |  | Target |
| choc | Konsum von Schoko-lade |  | Target |
| coke | Konsum von Kokain |  | Target |
| crack | Konsum von Crack |  | Target |
| ecstasy | Konsum von Ecstacy |  | Target |
| heroin | Konsum von Heroin |  | Target |
| ketamine | Konsum von Ketamin |  | Target |
| legalh | Konsum von Legal Highs |  | Target |
| lsd | Konsum von LSD |  | Target |
| meth | Konsum |  | Target |
| mushrooms | Konsum von Pilzen |  | Target |
| nicotine | Konsum von Nikotin |  | Target |
| semer |  |  | Target |
| vsa |  |  | Target |

## 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. Zudem sind 19 Zielvariablen in Form von einzelnen Drogenarten enthalten.

Da jedoch das Trainieren der jeweiligen Machine Learning Modellen 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.

## Lage- und Streuungsmaße

## Datenvisualisierung

# Data Preparation

## Fehlwerte

## Duplikate

## Weitere Kriterien für Datenqualität

## Skalierung und Standardisierung

1. <https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified> [↑](#footnote-ref-1)
2. <https://arxiv.org/pdf/1506.06297v1> [↑](#footnote-ref-2)