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.

Anhand der im folgenden Kapitel aufgestellten Ziel-Formulierung

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)“. Er enthält Informationen über den Konsum verschiedener Drogen in Abhängigkeit von soziodemografischen und psychologischen Merkmalen.

## SMART-Ziel Formulierung

Bis zum Abschluss des 4. Semesters werden verschiedene Machine-Learning-Modelle (Random Forest, Entscheidungsbäume, SVM und Lineare Regression) systematisch verglichen, um Vorhersagen über die Wahrscheinlichkeit und Intensität des zukünftigen Konsums von Kokain, Cannabis und Nikotin auf Basis von Umfragen zu Persönlichkeitsmerkmalen zu treffen. Der Erfolg der Modelle wird anhand von Trainings- und Testfehlern, F1-Score und Accuracy quantifiziert, um das leistungsfähigste Modell für präzise Prognosen zu identifizieren.

## Literaturrecherche

# Data Understanding

## Übersichtstabelle Datensatz

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Beschreibung** | **Skalenniveau** | **Typ** |
| id | Eindeutige Identifikationsnummer | Nominal | Integer |
| age | Alter der Person | Ordinal | Feature |
| gender | Geschlecht | Nominal | Feature |
| education | Bildung | Ordinal | Feature |
| country | Herkunftsland | Nominal | Feature |
| ethnicity | Ethnische Zugehörigkeit | Nominal | Feature |
| nscore | Neutotizismus-Wert | Ordinal | Feature |
| escore | Extraversion-Wert | Ordinal | Feature |
| oscore | Offenheit-Wert | Ordinal | Feature |
| ascore | Verträglichkeit-Wert | Ordinal | Feature |
| cscore | Gewissenhaftigkeit-Wert | Ordinal | Feature |
| impuslive | Impulisivitätsskala | Ordinal | Feature |
| ss | Sensation Seeking | Ordinal | Feature |
| alcohol | Konsum von Alkohol | Ordinal | Target |
| amphet | Konsum von Amphetaminen | Ordinal | Target |
| amyl | Konsum von Amylnitriten | Ordinal | Target |
| benzos | Konsum von Benzodiazepinen | Ordinal | Target |
| caff | Konsum von Kaffein | Ordinal | Target |
| cannabis | Konsum von Cannabis | Ordinal | Target |
| choc | Konsum von Schokolade | Ordinal | Target |
| coke | Konsum von Kokain | Ordinal | Target |
| crack | Konsum von Crack | Ordinal | Target |
| ecstasy | Konsum von Ecstasy | 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 von Methamphetamin | Ordinal | Target |
| mushrooms | Konsum von p. Pilzen | Ordinal | Target |
| nicotine | Konsum von Nikotin | Ordinal | Target |
| semer | Konsum von fiktiver Droge | Ordinal | Target |
| vsa |  |  | Target |

## Relevante Merkmale

## Lage- und Streuungsmaße

## Datenvisualisierung

# Data Preparation

## Fehlwerte

## Duplikate

## Weitere Kriterien für Datenqualität

## Skalierung und Standardisierung