Data Exploration Project

wwi23b4 | 06.04.2025

prediction of drug consumption

salih kelmendi, Luis Rastetter

2025

# Inhaltsverzeichnis

[Inhaltsverzeichnis 1](#_Toc193789685)

[1. Vorwort 2](#_Toc193789686)

[2. Business Understanding 2](#_Toc193789687)

[2.1 Auswahl des Datensatzes 2](#_Toc193789688)

[2.2 Literaturrecherche 2](#_Toc193789689)

[2.3 SMART-Ziel Formulierung 3](#_Toc193789690)

[3. Data Understanding 4](#_Toc193789691)

[3.1 Übersichtstabelle Datensatz 4](#_Toc193789692)

[3.2 Relevante Merkmale 5](#_Toc193789693)

[3.3 Lage- und Streuungsmaße 5](#_Toc193789694)

[3.4 Datenvisualisierung 7](#_Toc193789695)

[4. Data Preparation 8](#_Toc193789696)

[4.1 Fehlwerte 8](#_Toc193789697)

[4.2 Duplikate 8](#_Toc193789698)

[4.3 Weitere Kriterien für Datenqualität 8](#_Toc193789699)

[4.4 Skalierung und Standardisierung 8](#_Toc193789700)

[5. Modelling 9](#_Toc193789701)

[6. Evaluation 9](#_Toc193789702)

[7. Ausblick 9](#_Toc193789703)

<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 werden 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. Diese sind für das Trainieren geeigneter Modelle alle relevant, weshalb keine der Features aus dem Datensatz entfernt werden.   
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.

## 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 mehr notwendig ist und wird somit für den weiteren Verlauf des Trainierens der Modelle beibehalten. Ebenfalls ist eine gleichmäßige Verteilung der Daten um einen zentralen Wert gegeben, sodass keine allzu robusten Modelle von Nöten sind, die eine besondere Variabilität berücksichtigen.

Die folgenden Abbildungen bestätigen diese Erkenntnisse. 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).

Ein Bild, das Text, Screenshot, Karte Menü, Schrift enthält.

Automatisch generierte Beschreibung

Abbildung 1: Lage- und Streuungsmaße der Features[[3]](#footnote-3)

Ein Bild, das Diagramm, Origami enthält.

Automatisch generierte Beschreibung

Abbildung 2: Violinplot der Features[[4]](#footnote-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 (Abbildung 3).

Ein Bild, das Diagramm, Text, Reihe enthält.

Automatisch generierte Beschreibung

Abbildung 3: Histogramme der Features[[5]](#footnote-5)

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, welche eine Einnahme jener Droge am Tag vor der Befragung darstellt.

Zur Verdeutlichung werden beispielhaft hierfür Cannabis (Abbildung 4) und Nikotin (Abbildung 5) verwendet.

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

Abbildung 4: Verteilung des Konsums von Cannabis[[6]](#footnote-6)

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

Abbildung 5: Verteilung des Konsums von Nikotin[[7]](#footnote-7)

# 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 das spätere Erstellen der Modelle ermöglicht.

## 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.

## 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.

## Skalierung und Standardisierung

# Modelling

# Evaluation

# Ausblick

1. <https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified> [↑](#footnote-ref-1)
2. <https://arxiv.org/pdf/1506.06297v1> [↑](#footnote-ref-2)
3. Eigene Abbildung [↑](#footnote-ref-3)
4. Eigene Abbildung [↑](#footnote-ref-4)
5. Eigene Abbildung [↑](#footnote-ref-5)
6. Eigene Abbildung [↑](#footnote-ref-6)
7. Eigene Abbildung [↑](#footnote-ref-7)