Drug Consumption

Nhat Bui, Johan Ferreira, Thilo Holstein

2025-03-06

Contents

1 Introduction

Drug use is a significant risk behavior with serious health consequences for individuals and society. Multiple factors contribute to initial drug use, including psychological, social, individual, environmental, and economic elements, as well as personality traits. While legal substances like sugar, alcohol, and tobacco cause more premature deaths, illegal recreational drugs still create substantial social and personal problems.

In this data science project, we aim to identify factors and patterns potentially explaining drug use behaviors through machine learning techniques. By analyzing demographic, psychological, and social variables in our dataset, we'll aim to uncover potential predictors, use machine learning methods to understand the complex relationships surrounding drug consumption, demonstrating how machine learning can reveal insights into behavioral patterns. While our findings won't directly inform interventions, this project showcases how data-driven approaches can enhance our understanding of complex social phenomena and provide valuable practice in applying machine learning to real-world datasets.

The database contains records for 1,885 respondents with 12 attributes including personality measurements (NEO-FFI-R, BIS-11, ImpSS), demographics (education, age, gender, country, ethnicity), and self-reported usage of 18 drugs plus one fictitious drug (Semeron). Drug use is classified into seven categories ranging from "Never Used" to "Used in Last Day." All input attributes are quantified as real values, creating 18 distinct classification problems corresponding to each drug. A detailed description of the variables can be found in the Column Decsription text file.

2 Cleaning and formatting the dataset

```
# Load required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(reshape2)
```

```
# Read the CSV file
drug_data <- read.csv("Data/drug_consumption.csv")
```

2.1 Fomatting the dataset

The original data set had all the values for most of the variables set to a random floating number representing a specific categorical value, we believe this was done in order to remove bias from the dataset. As the requirements of this project is different form the data set's original intention we had to replace these values with the original values in order to complete all the required steps for our project.

```r
################################################################################
# Define mappings and column information
################################################################################

# Column names for the dataset
column_names <- c(
  "Index", "ID", "Age", "Gender", "Education", "Country", "Ethnicity",
  "Nscore", "Escore", "Oscore", "Ascore", "Cscore", "Impulsive", "SS",
  "Alcohol", "Amphet", "Amyl", "Benzos", "Caff", "Cannabis", "Choc", "Coke",
  "Crack", "Ecstasy", "Heroin", "Ketamine", "Legalh", "LSD", "Meth",
  "Mushrooms", "Nicotine", "Semer", "VSA"
)

# Drug column names
drug_columns <- c(
  "Alcohol", "Amphet", "Amyl", "Benzos", "Caff", "Cannabis",
  "Choc", "Coke", "Crack", "Ecstasy", "Heroin", "Ketamine",
  "Legalh", "LSD", "Meth", "Mushrooms", "Nicotine", "Semer", "VSA"
)

# Mapping of drug consumption classes to their meanings
consumption_mapping <- c(
  "CL0" = "Never Used",
  "CL1" = "Used over a Decade Ago",
  "CL2" = "Used in Last Decade",
  "CL3" = "Used in Last Year",
  "CL4" = "Used in Last Month",
  "CL5" = "Used in Last Week",
  "CL6" = "Used in Last Day"
)

# Map Age values to their meaning
age_mapping <- c(
  "-0.95197" = "18-24",
  "-0.07854" = "25-34",
  "0.49788" = "35-44",
  "1.09449" = "45-54",
  "1.82213" = "55-64",
  "2.59171" = "65+"
)

# Map Gender values to their meaning
gender_mapping <- c(
  "0.48246" = "Female",
  "-0.48246" = "Male"
)

# Map Education values to their meaning
education_mapping <- c(
  "-2.43591" = "Left school before 16 years",
  "-1.73790" = "Left school at 16 years",
  "-1.43719" = "Left school at 17 years",
  "-1.22751" = "Left school at 18 years",
```

```
  "-0.61113" = "Some college or university, no certificate or degree",
  "-0.05921" = "Professional certificate/diploma",
  "0.45468" = "University degree",
  "1.16365" = "Masters degree",
  "1.98437" = "Doctorate degree"
)

# Map Country values to their meaning
country_mapping <- c(
  "-0.09765" = "Australia",
  "0.24923" = "Canada",
  "-0.46841" = "New Zealand",
  "-0.28519" = "Other",
  "0.21128" = "Republic of Ireland",
  "0.96082" = "UK",
  "-0.57009" = "USA"
)

# Map Ethnicity values to their meaning
ethnicity_mapping <- c(
  "-0.50212" = "Asian",
  "-1.10702" = "Black",
  "1.90725" = "Mixed-Black/Asian",
  "0.12600" = "Mixed-White/Asian",
  "-0.22166" = "Mixed-White/Black",
  "0.11440" = "Other",
  "-0.31685" = "White"
)
```

```
################################################################################
# Data Processing
################################################################################

# Rename the columns
colnames(drug_data) <- column_names

# Convert demographic columns to descriptive values
drug_data$Age <- age_mapping[as.character(drug_data$Age)]
drug_data$Gender <- gender_mapping[as.character(drug_data$Gender)]
drug_data$Education <- education_mapping[as.character(drug_data$Education)]
drug_data$Country <- country_mapping[as.character(drug_data$Country)]
drug_data$Ethnicity <- ethnicity_mapping[as.character(drug_data$Ethnicity)]

# Convert all drug consumption columns to descriptive values
for (col in drug_columns) {
  drug_data[[col]] <- consumption_mapping[as.character(drug_data[[col]])]
}
```

2.2 Investigating missing values

```
################################################################################
# Data Cleaning - Missing values
```

```
##############################################################################

# Remove unnecessary column
drug_data <- drug_data[, -which(names(drug_data) == "ID")]

# Check for NA values in each column
na_by_column <- sapply(drug_data, function(x) sum(is.na(x)))
cat("NA values by column:\n")
```

```
## NA values by column:
```

```
# Print only columns with NA values
print(na_by_column[na_by_column > 0])
```

```
## Education Ethnicity
##        99        83
```

```
cat("\n")
```

Only two columns contain missing values, affecting approximately 5% of the 1885 observations. Given the nature of these variables and the completeness of the rest of the data, we assume participants deliberately withheld this information. Therefore, we replaced the missing values with "Not Provided", allowing us to treat these instances as a distinct category.

```
# Replace NA values with "Not Provided"
drug_data$Education[is.na(drug_data$Education)] <- "Not Provided"
drug_data$Ethnicity[is.na(drug_data$Ethnicity)] <- "Not Provided"

# Save the updated dataframe back to CSV
write.csv(drug_data, "Data/cleaned.csv")
```

2.3 Investigating outliers

```
##############################################################################
# Data Cleaning - Looking for outliers
##############################################################################
# Define numeric columns for outlier analysis
numeric_cols <- c("Nscore", "Escore", "Oscore", "Ascore", "Cscore", "Impulsive", "SS")

# Function to identify outliers using IQR method
identify_outliers_iqr <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr

  return(data.frame(
    min = min(x, na.rm = TRUE),
```

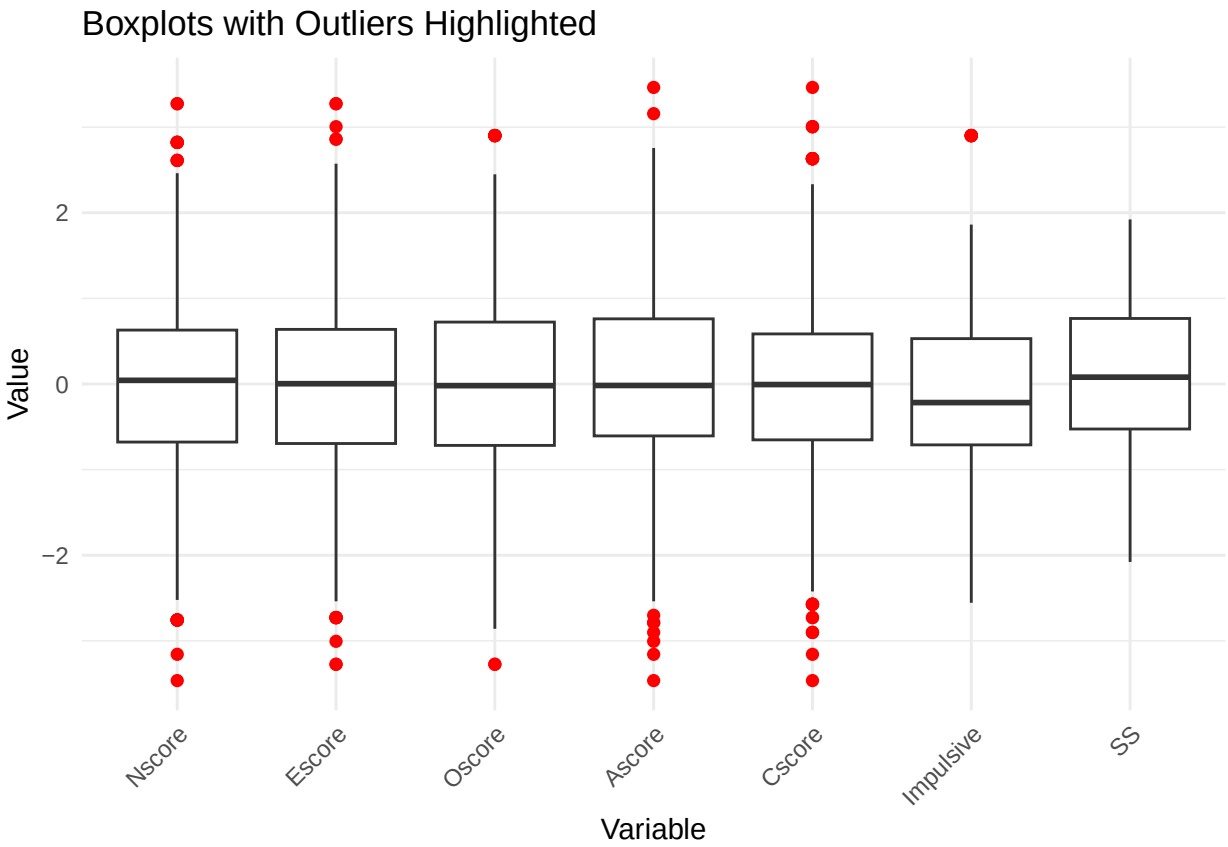
```
    q1 = q1,
    median = median(x, na.rm = TRUE),
    mean = mean(x, na.rm = TRUE),
    q3 = q3,
    max = max(x, na.rm = TRUE),
    iqr = iqr,
    lower_bound = lower_bound,
    upper_bound = upper_bound,
    n_outliers_below = sum(x < lower_bound, na.rm = TRUE),
    n_outliers_above = sum(x > upper_bound, na.rm = TRUE),
    total_outliers = sum(x < lower_bound | x > upper_bound, na.rm = TRUE),
    outlier_percentage = round(100 * sum(x < lower_bound | x > upper_bound, na.rm = TRUE) / length(x[!is
  ))
}

# Apply outlier detection to all numeric columns
outlier_summary <- data.frame()
for (col in numeric_cols) {
  result <- identify_outliers_iqr(drug_data[[col]])
  result$variable <- col
  outlier_summary <- rbind(outlier_summary, result)
}

# Create a function to visualize outliers with boxplots
plot_outliers <- function(drug_data, columns) {
  # Explicitly use reshape2::melt to avoid namespace issues
  melted_data <- reshape2::melt(drug_data[, columns], id.vars = NULL)

  # Create boxplot
  ggplot(melted_data, aes(x = variable, y = value)) +
    geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
    theme_minimal() +
    labs(title = "Boxplots with Outliers Highlighted",
         x = "Variable",
         y = "Value") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

# Visualize outliers
plot_outliers(drug_data, numeric_cols)
```



As can be seen from the box plots our data set has some values that are outside of the upper and lower bounds. All thought these values are technically outliers they are not extreme, still fall inside of the range of our expected values and conforms to a normal distribution.

3 Source

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