# Effects of Alcohol on Students Performance

**Name** : Arun Kumar Korra
**Email** : arunkumar.korraa@gmail.com
**Institution** : University College Of Engineering,Osmania University
**Submitted to** : CSR Box, IBM Skills Build
**Submitted date** : 3rd November, 2024

## Table of Contents

## Abstract

This project examines the effects of alcohol consumption on the academic performance of students at Stellenbosch University, with a focus on aligning findings with the Sustainable Development Goals (SDGs) 3 and 4, which address health and quality education. Using a dataset collected in 2023, the study analyzes various demographic and lifestyle factors, particularly alcohol intake, to develop a predictive model of student GPA.

The methodology involves data preprocessing, encoding categorical variables, and employing Support Vector Regression (SVR) for model development. The results indicate a negative correlation between alcohol consumption and GPA, with students who consume more alcohol typically achieving lower academic performance. Additionally, students who reported skipping classes due to alcohol use displayed significantly reduced GPAs.

Visualizations, including scatter plots, box plots, and heatmaps, effectively illustrate these findings and emphasize the need for promoting healthier lifestyles within university settings. The study concludes that addressing alcohol consumption is vital for enhancing student well-being and academic success, suggesting avenues for future research to identify effective interventions.

## 1. Introduction

### Background:

The Sustainable Development Goals (SDGs) aim to tackle important global issues, particularly in health and education. SDG 3 focuses on promoting well-being, while SDG 4 emphasizes the need for quality education. These goals are especially relevant in university settings, where student lifestyles, including alcohol consumption, can greatly affect academic performance.

### Problem Statement

This project investigates how alcohol consumption impacts student performance. By analyzing various lifestyle and demographic factors, the study aims to provide insights that can help improve student well-being and academic success.

## 2. Objective

The main objective of this project is to create a predictive model that identifies how different factors, including alcohol consumption, affect the academic performance of students at Stellenbosch University. This research supports the goals of SDGs 3 and 4 by enhancing our understanding of student health and education.

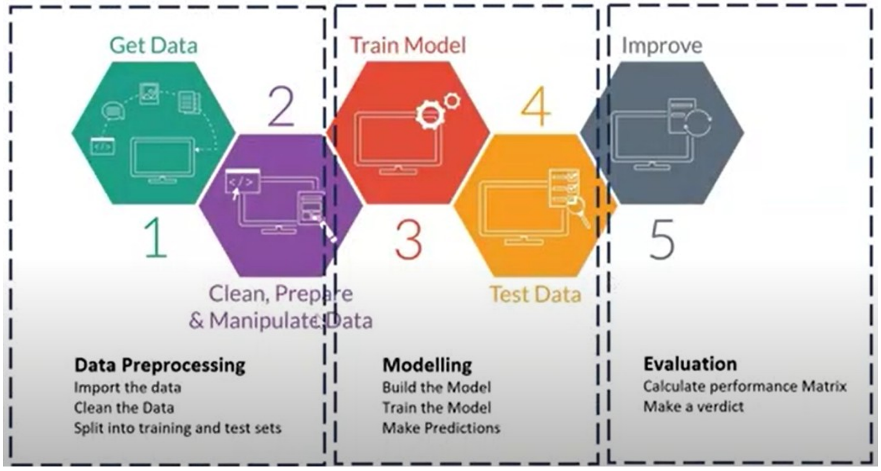
## 3. Focus on SDGs

This project targets both SDG 3: Good Health and Well-Being and SDG 4: Quality Education. It emphasizes the importance of promoting healthy lifestyles among students to enhance their academic performance, thereby contributing to their overall well-being and educational outcomes.

## 4. Sources

1. Data is collected from the Kaggle.
2. Student survey data collected in 2023.
3. Academic performance records from Stellenbosch University.
4. Literature on the effects of lifestyle choices on student performance.

## 5. Methodology



a. **Data Collection:**

The data for this study, titled **"Effects of Alcohol on Student Performance"** was collected by **Joshua Naude**. The study was conducted within the **Department of Statistics and Actuarial Science** at **Stellenbosch University, South Africa**. The dataset includes various demographic and academic performance variables related to students, specifically focusing on their alcohol consumption behaviours and academic outcomes.

b. **Data Preprocessing**

The collected data was processed using Python libraries including Pandas, NumPy, and Scikit-learn. The preprocessing steps included:

- Loading of the data using the pandas.
- Renaming the columns for clarity, better understanding.
- Missing values were checked and removed.

c. **Data Encoding**

Categorical variables were encoded using LabelEncoder from Scikit-learn. This encoding process was applied to the following variables:

- Gender
- Number of drinks
- Socializing frequency
- Relationship status
- Parent approval

d. **Feature selection and Target variable**

The independent variables (features) selected for analysis included:

- Gender
- Number of drinks
- Skipped classes due to alcohol
- Parent approval
- Socializing frequency

The dependent variable (target) for this study was the students' GPA for the year 2023.

e. **Data splitting**

The dataset was split into training and testing sets using Scikit-learn's "train_test_split" function. The training set comprised 80% of the data, while the remaining 20% was reserved for testing the model's performance.

f. **Data Standardization**

To ensure that all features contributed equally to the model training, feature scaling was performed using StandardScaler. This process standardized the training and testing datasets.

g. **Model Development**

A Support Vector Regression (SVR) model was employed to predict the GPA based on the selected features. The model was configured to use a linear kernel for this analysis.

h. **Model Evaluation**

Model performance was assessed using:

- **Mean Squared Error (MSE)**: This metric quantified the average squared difference between actual and predicted GPAs.
- **$R^2$ Score**: This statistic indicated the proportion of variance in the GPA that could be explained by the model.

i. **Hyperparameter Tuning**

To enhance the model's performance, a grid search was conducted using ”GridSearchCV“, focusing on hyperparameters such as the regularization parameter (C) and the epsilon value for the SVR.

j. **Data Visualization**

Several visualizations were created to illustrate the relationships between variables:

- **Scatter Plots**: These plots compared actual vs. predicted GPAs for both training and testing datasets, helping to visualize model performance.
- **Heatmap**: A heatmap was generated to showcase the correlation between numerical features in the dataset.
- **Box and Bar Plots**: These plots were utilized to display GPA distributions across different drinking levels and socializing frequencies.

## 6. Findings

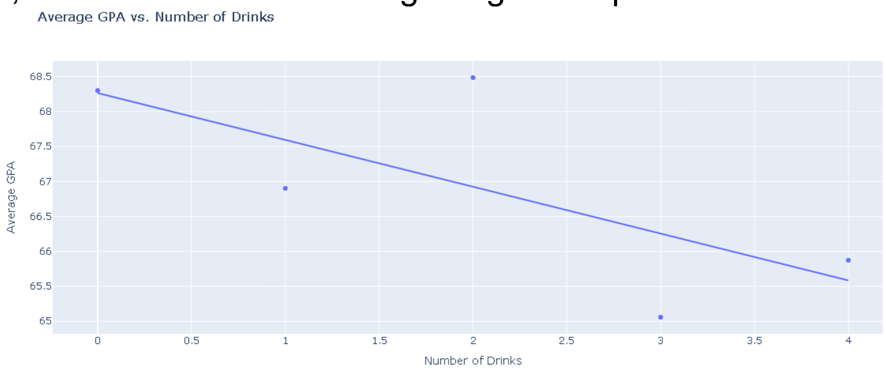
The analysis revealed several key insights regarding the impact of alcohol consumption on academic performance among students at Stellenbosch University:

1. **Correlation Between Alcohol Consumption and GPA**: The data indicated a negative correlation between the number of drinks consumed and students' GPAs. Higher alcohol consumption was associated with lower academic performance.
2. **Impact of Skipping Classes**: Students who reported skipping classes due to alcohol consumption tended to have significantly lower GPAs compared to those who did not. This finding underscores the potential consequences of alcohol-related absenteeism on academic success.
3. **Socializing Habits**: The study found that students who engaged in frequent socializing and partying had varied GPA outcomes, suggesting that not all social activities negatively impacted academic performance, but excessive drinking did.

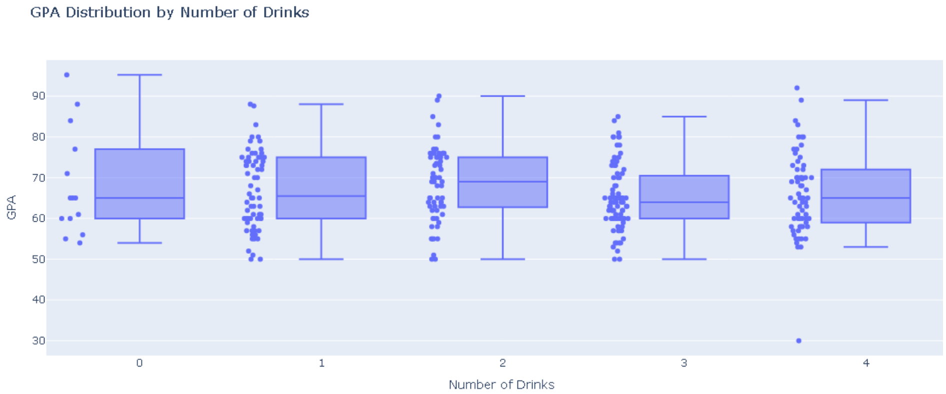
### Visualizations

Several visualizations were created to illustrate these findings:

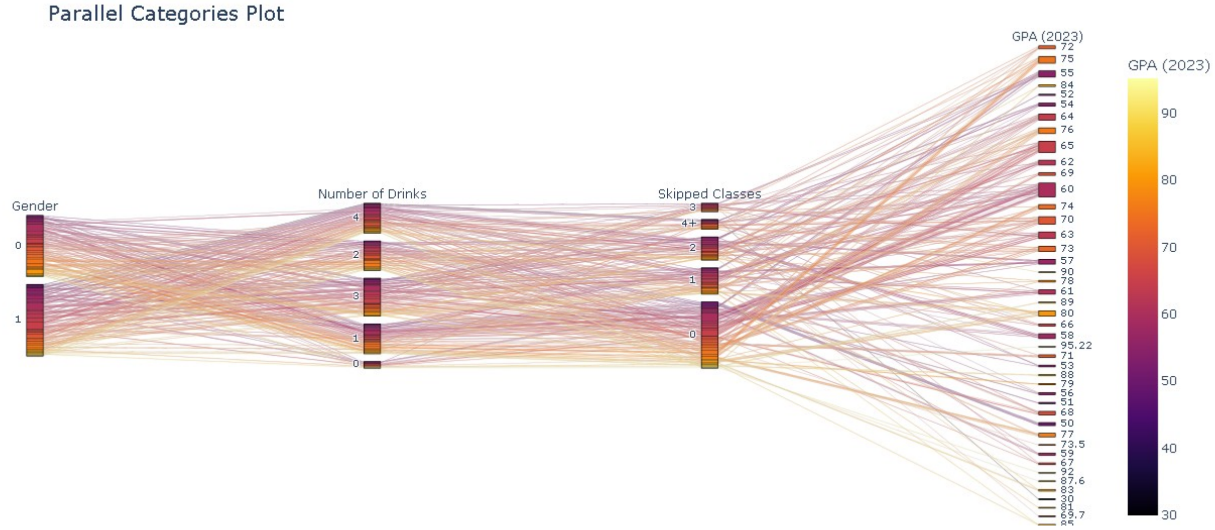
1. **Scatter Plot of GPA vs. Number of Drinks**:
   A scatter plot showing the relationship between the number of drinks consumed and the GPA. Points representing individual students will be plotted, with a trendline indicating the general pattern.



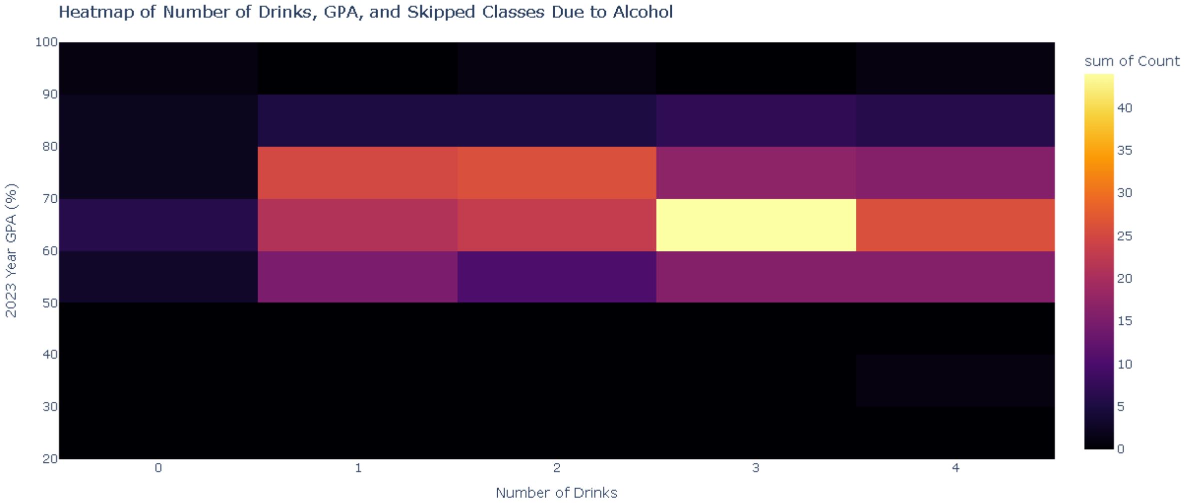
2. **Box Plot of GPA by Alcohol Consumption**:
   A box plot displaying the distribution of GPAs across different levels of alcohol consumption (e.g., none, low, moderate, high). This plot will help visualize the impact of varying drinking levels on academic performance.



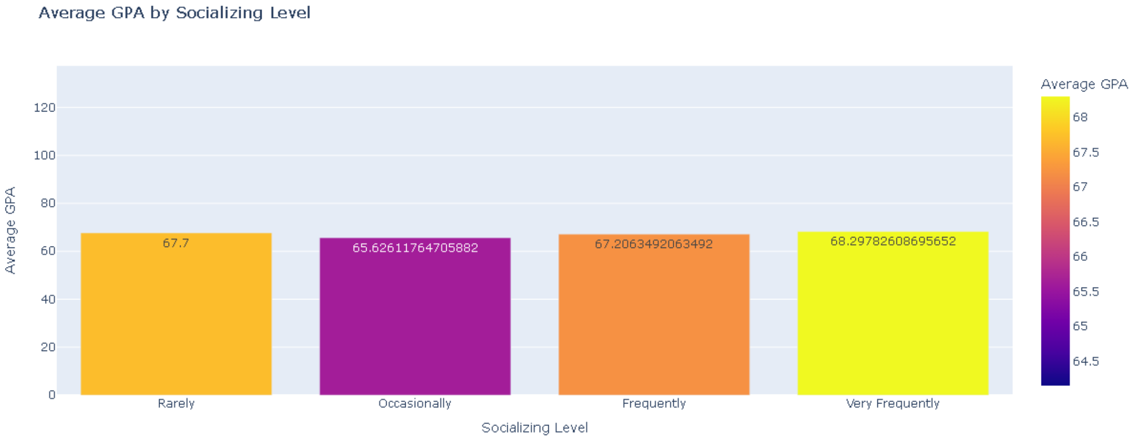
3. **Parallel Categories Plot**:
   A parallel categories plot visualizing the relationships between gender, number of drinks, skipped classes due to alcohol, and GPA. The colors represent GPA, providing context for how these factors interact.



4. **Heatmap of Number of Drinks, GPA, and Skipped Classes**:
   A heatmap showing the density of occurrences for the number of drinks, GPA, and instances of skipping classes due to alcohol. This visualization helps identify patterns in how these variables interact.



5. **Bar Plot of Average GPA by Socializing Frequency**:
   A bar plot comparing the average GPA of students based on their self-reported socializing frequency. This visualization conveys how social habits relate to academic performance.



## 7. Conclusion

This project investigated the impact of alcohol consumption on academic performance among students at Stellenbosch University, aligning with Sustainable Development Goals 3 and 4, which focus on health and quality education. The analysis revealed a negative correlation between alcohol intake and GPA, with students who drank more frequently achieving lower academic results. Additionally, those who skipped classes due to alcohol use faced significant declines in performance.

Visualizations highlighted these relationships, underscoring the need for promoting healthier lifestyles among students. The findings suggest that addressing alcohol consumption in university settings is crucial for enhancing student well-being and academic success. Future research could explore effective interventions to further reduce alcohol-related impacts on academic performance.

## 8. References

- Kaggle: **Effects of Alcohol on Student Performance**
- Scikit-learn Documentation
- Pandas Documentation
- Plotly.express Documentation
- Seaborn Documentation
- Matplotlib Documentation

## 9. Appendices

- **Appendix A:**
  A code snippet is used in this project

```
import numpy as np  # NumPy for numerical operations and linear algebra
import pandas as pd  # Pandas for data processing and handling CSV files
import seaborn as sns  # Seaborn for statistical data visualization
import matplotlib.pyplot as plt  # Matplotlib for plotting
import plotly.express as px  # Plotly for interactive data visualization
```

```
df = pd.read_csv("Stats survey.csv")

new_cols = {
    df.columns[1]: "Gender",  # Rename the second column to 'Gender'
    df.columns[2]: "Overall GPA, %",  # Rename the third column to 'Overall GPA, %'
    df.columns[3]: "Year",  # Rename the fourth column to 'Year'
    df.columns[4]: "Faculty",  # Rename the fifth column to 'Faculty'
    df.columns[5]: "2023 year GPA, %",  # Rename the sixth column to '2023 year GPA, %'
    df.columns[6]: "Accommodation 2023",  # Rename the seventh column to 'Accommodation
2023'
    df.columns[8]: "Scholarship Student",  # Rename the ninth column to 'Scholarship
Student'
    df.columns[9]: "Additional Study hours",  # Rename the tenth column to 'Additional
Study hours'
    df.columns[10]: "Socialising, partying, drinking",  # Rename the eleventh column to
'Socialising, partying, drinking'
    df.columns[11]: "Number of drinks",  # Rename the twelfth column to 'Number of
drinks'
    df.columns[12]: "Skipped classes because of alcohol",  # Rename the thirteenth
column to 'Skipped classes because of alcohol'
    df.columns[13]: "Failed modules",  # Rename the fourteenth column to 'Failed
modules'
    df.columns[14]: "Relationship",  # Rename the fifteenth column to 'Relationship'
    df.columns[15]: "Parent approval",  # Rename the sixteenth column to 'Parent
approval'
    df.columns[16]: "Bonding with parents"  # Rename the seventeenth column to 'Bonding
with parents'
}

# Renaming the DataFrame columns using the new_cols dictionary
df.rename(columns=new_cols, inplace=True)  # Apply the new column names to the DataFrame

df.isnull().sum()  # Check for missing values

df.dropna(inplace=True)  # Remove rows with missing values

from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder
le = LabelEncoder()

# Apply LabelEncoder to the specified columns
df['Gender'] = le.fit_transform(df['Gender'])
df['Number of drinks'] = le.fit_transform(df['Number of drinks'])
df['Socialising, partying, drinking'] = le.fit_transform(df['Socialising, partying,
drinking'])
df['Relationship'] = le.fit_transform(df['Relationship'])
df['Parent approval'] = le.fit_transform(df['Parent approval'])

x = df[['Gender', 'Number of drinks', 'Skipped classes because of alcohol',"Parent
approval" ,"Socialising, partying, drinking" ]]
y = df['2023 year GPA, %']  # or 'Overall GPA'

# Change 'Number of drinks' to numbers; if there's a problem, use 4 instead of missing
values
x['Number of drinks'] = pd.to_numeric(x['Number of drinks'], errors='coerce').fillna(4)

# Change 'Skipped classes because of alcohol' to numbers; if there's a problem, use 4
instead of missing values
x['Skipped classes because of alcohol'] = pd.to_numeric(x['Skipped classes because of
alcohol'], errors='coerce').fillna(4)

# Change 'Parent approval' to numbers; if there's a problem, use 4 instead of missing
values
x['Parent approval'] = pd.to_numeric(x['Parent approval'], errors='coerce').fillna(4)

from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,
random_state=42)

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

from sklearn.svm import SVR

model = SVR(kernel='linear')  # You can experiment with different kernels (linear, rbf,
etc.)
model.fit(x_train, y_train)

from sklearn.metrics import mean_squared_error, r2_score

y_pred = model.predict(x_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R² Score: {r2}')

from sklearn.model_selection import GridSearchCV

param_grid = {'C': [0.1, 1, 10], 'epsilon': [0.1, 0.2, 0.5]}
grid = GridSearchCV(SVR(), param_grid, refit=True)
grid.fit(x_train, y_train)

# Visualizing the Training set results
plt.figure(figsize=(10, 6))

# Scatter plot of actual GPAs in the training set
plt.scatter(x_train[:, 1], y_train, color='red', label='Actual GPA (Train)', alpha=0.6)

# Scatter plot of predicted GPAs in the training set
plt.scatter(x_train[:, 1], model.predict(x_train), color='blue', label='Predicted GPA
(Train)', alpha=0.6)

# Add titles and labels
plt.title('Actual vs Predicted GPA on Training Set')
plt.xlabel('Number of Drinks')
plt.ylabel('GPA')
plt.legend()

# Show the plot
plt.show()

# Visualizing the Test set results
plt.figure(figsize=(10, 6))

# Scatter plot of actual GPAs in the test set
plt.scatter(x_test[:, 1], y_test, color='green', label='Actual GPA (Test)', alpha=0.6)

# Scatter plot of predicted GPAs in the test set
plt.scatter(x_test[:, 1], y_pred, color='orange', label='Predicted GPA (Test)',
alpha=0.6)

# Add titles and labels
plt.title('Actual vs Predicted GPA on Test Set')
plt.xlabel('Number of Drinks')
plt.ylabel('GPA')
plt.legend()

# Show the plot
plt.show()


# Calculate the average GPA based on the number of drinks
```

```
avg_gpa_drinks = df.groupby('Number of drinks')['2023 year GPA, %'].mean().reset_index()

# Scatter plot to show relationship between Number of drinks and GPA
fig_scatter = px.scatter(
    avg_gpa_drinks,
    x='Number of drinks',
    y='2023 year GPA, %',
    title='Average GPA vs. Number of Drinks',
    labels={'Number of drinks': 'Number of Drinks', '2023 year GPA, %': 'Average GPA'},
    trendline='ols'  # Adding a trendline
)
fig_scatter.show()

# Heatmap to visualize correlations between numeric features
numeric_df = df.select_dtypes(include=[np.number])  # Select only numeric columns
correlation_matrix = numeric_df.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Correlation Heatmap')
plt.show()

# Box plot to visualize the distribution of GPA across different levels of drinking
fig_box = px.box(
    df,
    x='Number of drinks',
    y='2023 year GPA, %',
    title='GPA Distribution by Number of Drinks',
    labels={'Number of drinks': 'Number of Drinks', '2023 year GPA, %': 'GPA'},
    points='all'  # Show all points
)
fig_box.show()

# Bar plot to compare average GPA across different socializing levels
# Create a mapping for socializing levels
socializing_labels = {
    0: 'Rarely',
    1: 'Occasionally',
    2: 'Frequently',
    3: 'Very Frequently'
}

# Assuming socializing levels are encoded as integers
avg_gpa_socializing = df.groupby('Socialising, partying, drinking')['2023 year GPA,
%'].mean().reset_index()

# Replace the encoded values with descriptive names
avg_gpa_socializing['Socialising, partying, drinking'] =
avg_gpa_socializing['Socialising, partying, drinking'].map(socializing_labels)

fig_bar = px.bar(
    avg_gpa_socializing,
    x='Socialising, partying, drinking',
    y='2023 year GPA, %',
    title='Average GPA by Socializing Level',
    labels={'Socialising, partying, drinking': 'Socializing Level', '2023 year GPA, %':
'Average GPA'},
    color='2023 year GPA, %',
    text='2023 year GPA, %'  # Display GPA values on the bars
)
fig_bar.show()
```