DSA 210: Introduction to Data Science Fall 2025-2026

Alcohol and Tobacco Consumption Analysis

Prepared by: Ömer Berke Uzun - 34335

Table of Contents

1. Motivation

Alcohol and tobacco consumption is a prevalent issue in various countries throughout the world. This project aims to analyze and understand the socio-economic factors such as GDP, education levels, and population statistics contributing to the usage of these substances and whether the government's efforts to control them are effective. By comparing different countries' features, the analysis will hopefully help us understand how we can reduce these substances consumption rates.

The project will thus focus on these questions:

1. How do socio-economic factors correlate with the usage of alcohol and tobacco?
2. Does government control help reduce these substance's consumption in a meaningful way?
3. Can we predict the consumption rate of these substances and find factors that reduce them?

2. Data Sources

The datasets used to conduct this analysis are as follows:

1. Global Country Information Dataset 2023
   - Dataset to get the socio-economic factors of each country
   - Features used:
     - Country: Name of the country.
     - Density (P/Km2): Population density measured in persons per square kilometer.
     - Birth Rate: Number of births per 1,000 population per year.
     - CPI: Consumer Price Index, a measure of inflation and purchasing power.
     - CPI Change (%): Percentage change in the Consumer Price Index compared to the previous year.
     - GDP: Gross Domestic Product, the total value of goods and services produced in the country.
     - Gross Primary Education Enrollment (%): Gross enrollment ratio for primary education.

- Gross Tertiary Education Enrollment (%): Gross enrollment ratio for tertiary education.
- Life Expectancy: Average number of years a newborn is expected to live.
- Total Tax Rate: Overall tax burden as a percentage of commercial profits.
- Unemployment Rate: Percentage of the labor force that is unemployed.
- Urban Population: Percentage of the population living in urban areas.

2. Tobacco: current tobacco use, tobacco smoking and cigarette smoking, age-standardized

- Dataset to get the tobacco usage of each country
- Features used:
  - Country: Name of the country.
  - Estimate of current tobacco use prevalence (%) (age-standardized rate): Estimated usage rate of tobacco.

3. Alcohol, total per capita (15+) consumption (in litres of pure alcohol) (SDG Indicator 3.5.2)

- Dataset to get the alcohol usage of each country
- Features used:
  - Country: Name of the country.
  - Alcohol, total per capita (15+) consumption (in litres of pure alcohol) (SDG Indicator 3.5.2), three-year average: Average consumption of alcohol per capita.

4. Alcohol policy: adopted written national policy on alcohol

- Dataset to get the alcohol policies of each country
- Features used:
  - Country: Name of the country.
  - Adopted written national policy on alcohol: Does the country have a written policy to control alcohol usage.

5. Tobacco control: Monitor: national tobacco control programmes

- Dataset to get the tobacco policies of each country
- Features used:
  - Country: Name of the country.
  - Government objectives on tobacco control exist: Does the government try to control the tobacco usage.

3. Data Analysis

Data analysis was performed using merger.py where the datasets were collected and prepared and project.ipynb where the rest of the steps were performed.

3.1 Data Collection and Preparation

The datasets were first downloaded through their respective sites and then their features were filtered to retain the features mentioned in the previous section. Then they were all merged into a single dataframe and null values were cleaned. Afterwards Alcohol and Tobacco Regulation columns were cleaned of unwanted values to only have “Yes” or “No” values. Applicable columns were converted to from object type to numeric type to allow proper analysis. Finally, the dataframe was exported into merged_data.csv to begin data analysis.

3.2 Exploratory Data Analysis (EDA)

EDA is used to gain an intuitive understanding of the data by visualizing distributions along with heatmaps. This understanding helps prepare for deeper analysis through hypothesis tests and allows us to better interpret results.

3.2.1 Distribution and Relation Plots

The first distributions explored were of the target columns Alcohol Consumption and Tobacco Usage [1].

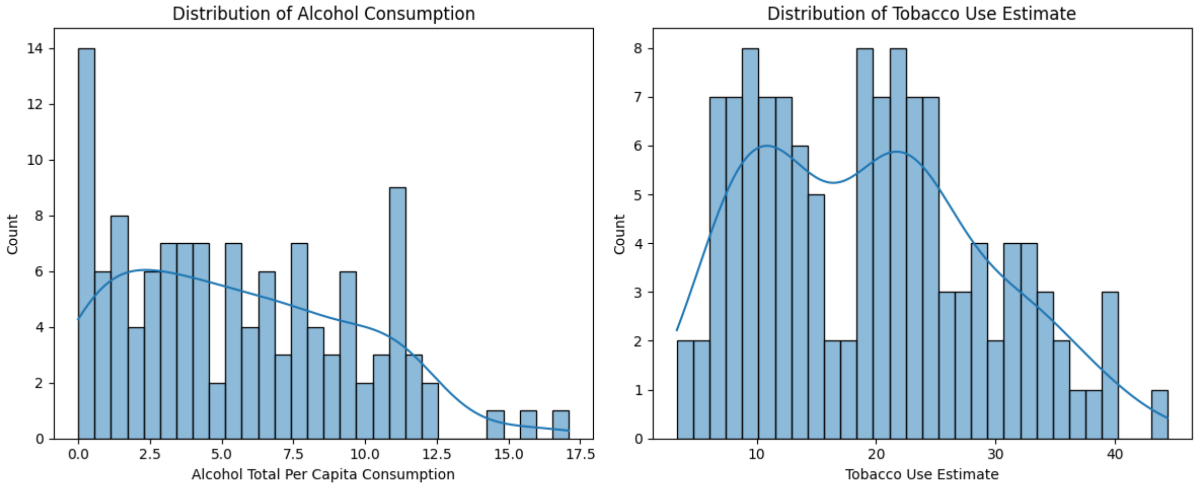


Image 1: Histogram charts of Alcohol Consumption and Tobacco Usage

The histograms help visualize the distribution of alcohol and tobacco consumption. It can be seen from the image that the alcohol consumption chart is right skewed which shows that most countries tend to drink relatively less alcohol while few with very high consumption increase the average. The tobacco use estimate chart is much closer to a symmetric shape with a broad plateau as compared to the alcohol consumption chart. This shape shows that tobacco use is more evenly distributed although most countries still have a smoker for every 5-10 people.

The second chart explores the total number of countries with alcohol and tobacco regulations in place [2].

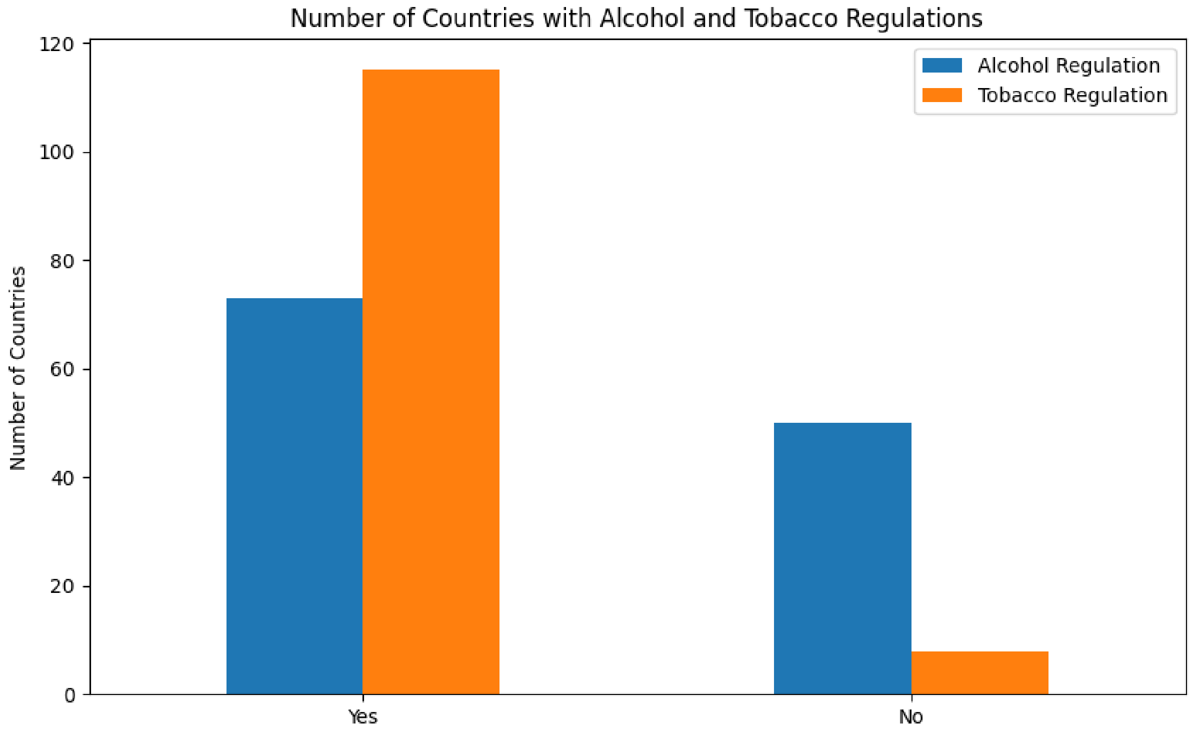


Image 2: Bar charts of Alcohol Regulation and Tobacco Regulation

The image shows that the number of countries that have alcohol regulations is slightly higher than the number of countries that do not. This indicates that alcohol is only moderately regulated with a sizeable number of countries lacking strong or formal regulations even when subnational policies were counted as regulations existing. Tobacco regulations on the other hand show a very high number of countries that have regulations. This shows that tobacco regulation is almost globalized with almost every country having some sort of regulation in place, much likely due to how tobacco is perceived by the public as compared to alcohol. These are also in line with the previous graphs showing uneven alcohol consumption between countries with outliers while tobacco usage showed a more even distribution which is likely due to the distribution of these regulations.

This leads us to the next plot where the regulations effect on the consumption of alcohol and tobacco was explored [3].

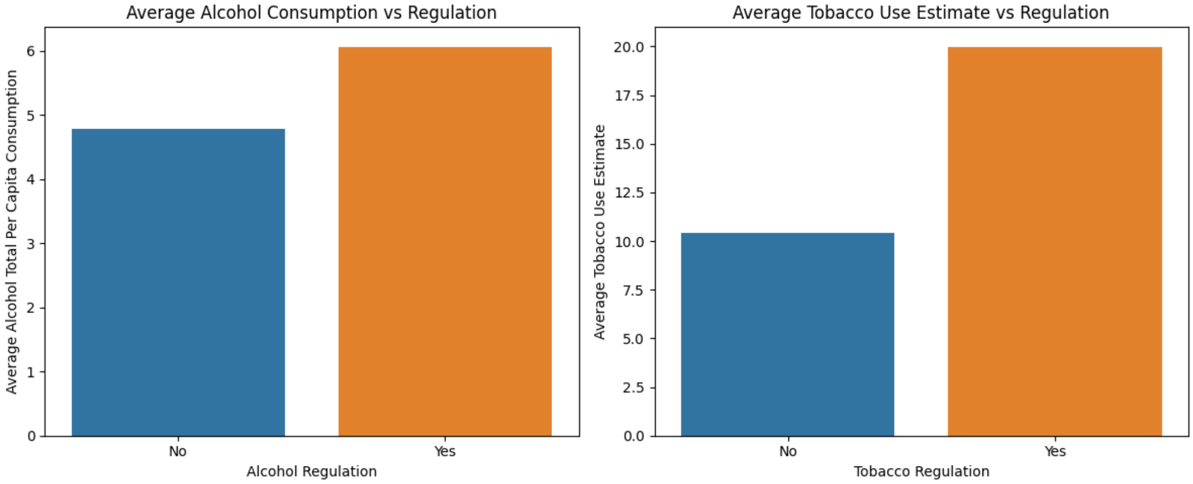


Image 3: Bar charts of Average Alcohol and Tobacco consumption for countries with and without regulations

The image showed a surprising result in regulations' effect on consumption. Seemingly, average consumption of both alcohol and tobacco are actually higher for countries with regulations in place. This suggests a reverse relationship between these factors where instead of regulations lowering the usage of these substances, they are actually implemented when the usage of these substances is high in the first place. This is a classic example of survivorship bias, as countries with low consumption of these substances do not need regulations in the first place, instead regulations start getting introduced when consumption reaches a level where health concerns start forming.

The next exploration was done to find out some features relationship with the target factors through the usage of scatterplots [4].

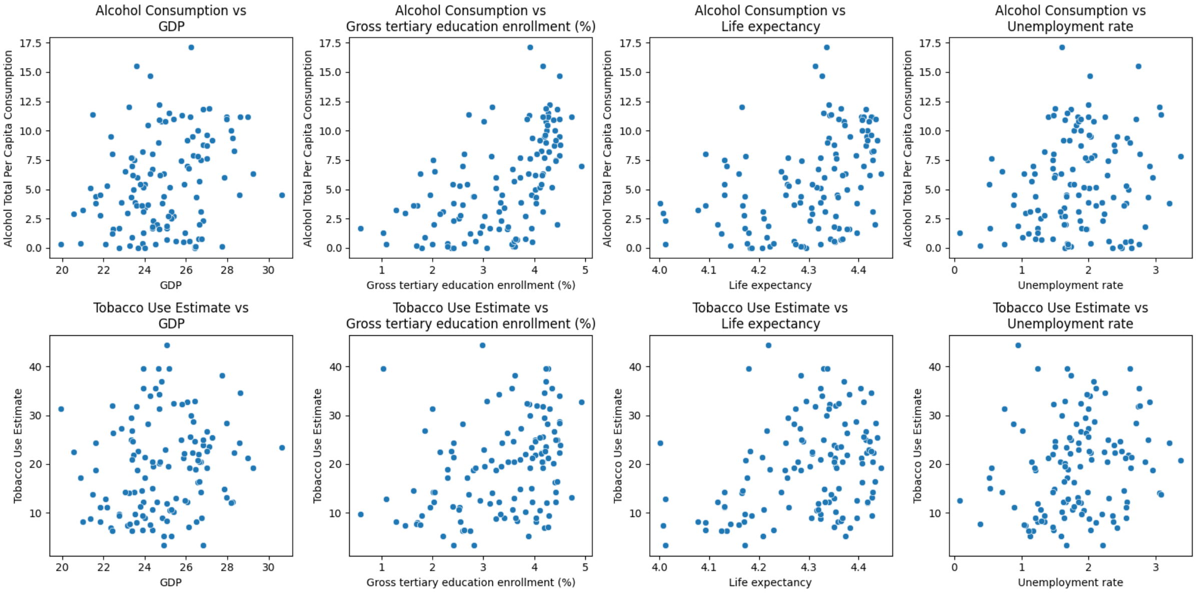


Image 4: Scatterplots showing relationship between Alcohol and Tobacco consumption and log of various features

Starting with GDP, the plots show that GDP has a low but seemingly positive correlation with alcohol consumption which is likely due to better affordability. Tobacco usage on the other hand has almost no correlation with GDP, but the graph does show a similar distribution to the previous tobacco distribution chart, likely due to high income countries having stricter regulations and low-income countries not having affordability.

For the relationships with gross tertiary education enrollment, highly educated countries seemingly tend to drink more alcohol but with large variability, likely due to drinking norms and education's correlation with income and urban lifestyle. A similar pattern is also observed for tobacco which shows that education itself does not eliminate the consumption of these substances.

Alcohol consumption also seemingly shows a positive trend with increased life expectancy which might seem counter-intuitive at first due to the health risks introduced by alcohol consumption. However, life expectancy is a sign of developed countries leading to higher consumption whose introduced health risks are likely mitigated by their advanced healthcare systems. The same relationship can also be seen for tobacco usage likely due to the same reasons, and it should also be noted that the harm caused by these substances has a lag and is chronic which allows high consumption to coexist with high life expectancy.

Unemployment rate on the other hand seemingly shows no correlation with alcohol consumption and tobacco usage most likely caused by it not being a good enough representation of the stress levels of the population by itself.

The final plots explored were boxplots and violinplots of alcohol consumption and tobacco usage to better understand their distributions [5].

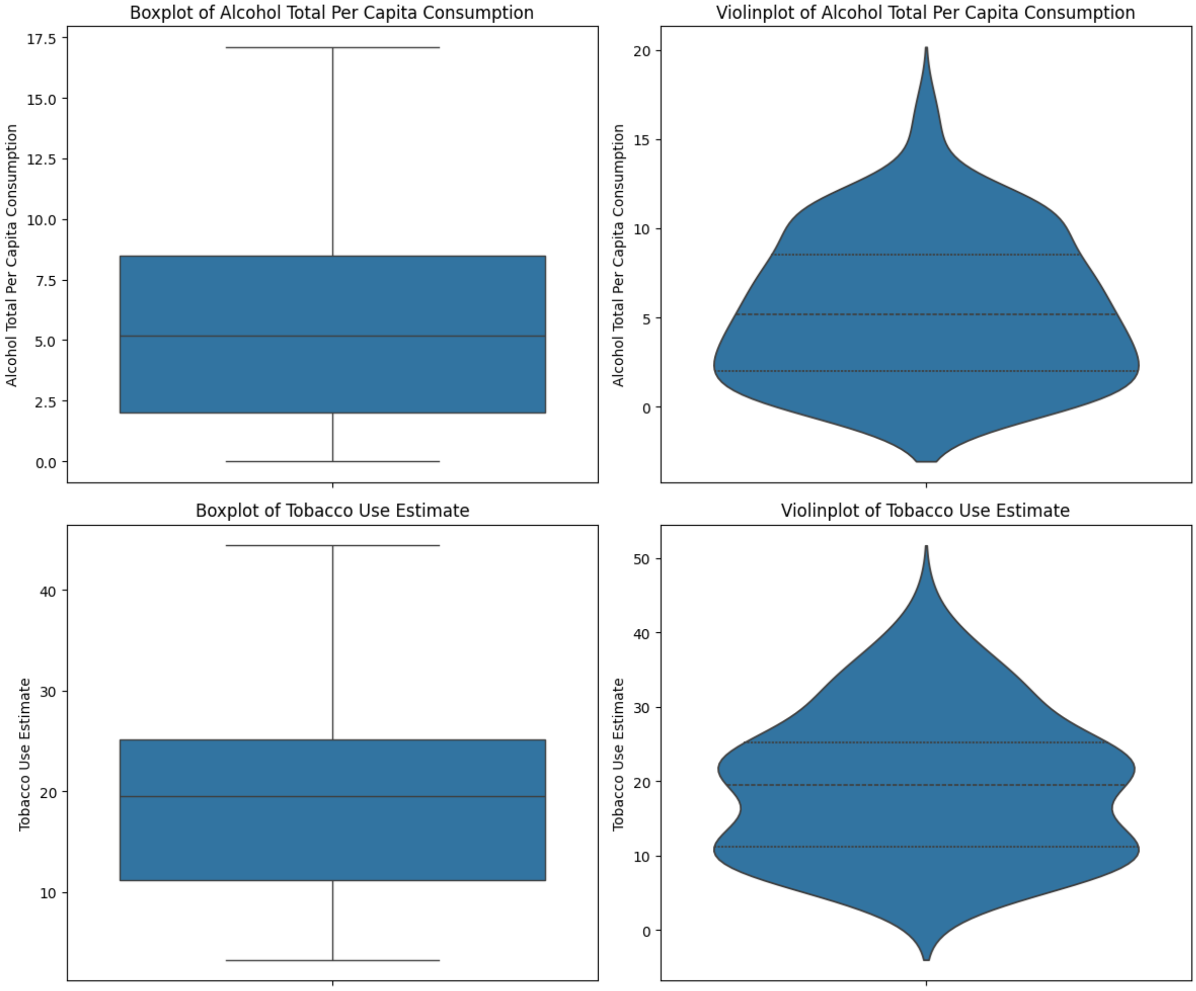


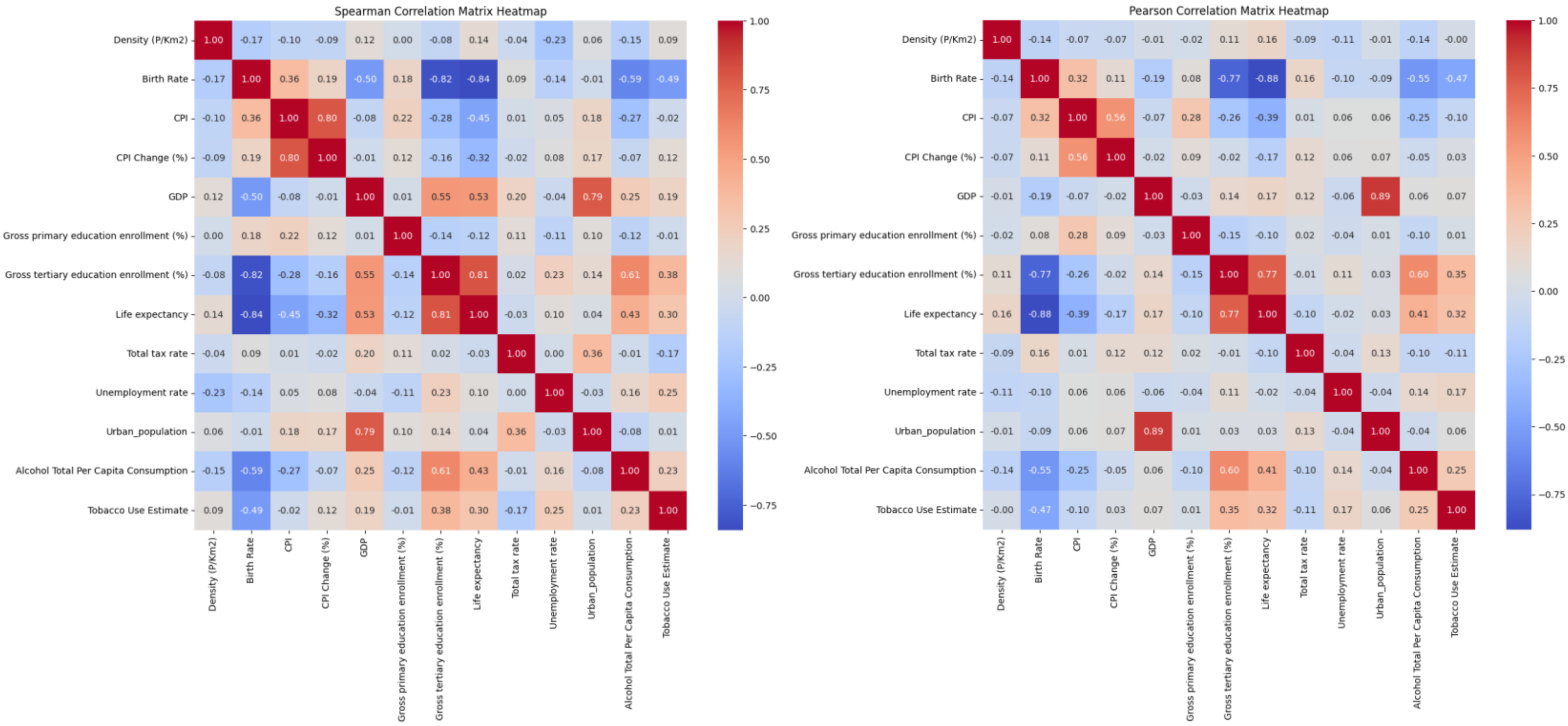
Image 5: Boxplots and Violinplots showing Alcohol Consumption and Tobacco Usage

The boxplot of alcohol consumption shows roughly 5 liters of pure alcohol consumption per capita, that is roughly 100 liters of beer per year on average, and it also shows a strong right skew with an upper whisker extending all the way to 17 liters. The violinplot shows high density at lower levels of consumption with a long and thin upper tail which is in line with the previous plots.

The boxplot of tobacco shows roughly 20% usage of tobacco on average which means every 1 in 5 people uses tobacco on average. The violinplot of tobacco provides similar findings to the boxplot with 2 density peaks and a wider upper tail as compared to alcohol which is once again in line with the previous distribution plots.

3.2.2 Correlation Heatmaps

Two correlation matrix heatmaps were plotted, one for Pearson correlation to explore linear relationships and one for Spearman correlation to be able to explore non-linear relationships as well [6,7].



	Density (P/Km2)	Birth Rate	CPI	CPI Change (%)	GDP	Gross primary education enrollment (%)	Gross tertiary education enrollment (%)	Life expectancy	Total tax rate	Unemployment rate	Urban_population	Alcohol Total Per Capita Consumption	Tobacco Use Estimate
Density (P/Km2)	1.00	-0.17	-0.10	-0.09	0.12	0.00	-0.08	0.14	-0.04	-0.23	0.06	-0.15	0.09
Birth Rate	-0.17	1.00	0.36	0.19	-0.50	0.18	-0.82	-0.84	0.09	-0.14	-0.01	-0.59	-0.49
CPI	-0.10	0.36	1.00	0.80	-0.08	0.22	-0.28	-0.45	0.01	0.05	0.18	-0.27	-0.02
CPI Change (%)	-0.09	0.19	0.80	1.00	-0.01	0.12	-0.16	-0.32	-0.02	0.08	0.17	-0.07	0.12
GDP	0.12	-0.50	-0.08	-0.01	1.00	0.01	0.55	0.53	0.20	-0.04	0.79	0.25	0.19
Gross primary education enrollment (%)	0.00	0.18	0.22	0.12	0.01	1.00	-0.14	-0.12	0.11	-0.11	0.10	-0.12	-0.01
Gross tertiary education enrollment (%)	-0.08	-0.82	-0.28	-0.16	0.55	-0.14	1.00	0.81	0.02	0.23	0.14	0.61	0.38
Life expectancy	0.14	-0.84	-0.45	-0.32	0.53	-0.12	0.81	1.00	-0.03	0.10	0.04	0.43	0.30
Total tax rate	-0.04	0.09	0.01	-0.02	0.20	0.11	0.02	-0.03	1.00	0.00	0.36	-0.01	-0.17
Unemployment rate	-0.23	-0.14	0.05	0.08	-0.04	-0.11	0.23	0.10	0.00	1.00	-0.03	0.16	0.25
Urban_population	0.06	-0.01	0.18	0.17	0.79	0.10	0.14	0.04	0.36	-0.03	1.00	-0.08	0.01
Alcohol Total Per Capita Consumption	-0.15	-0.59	-0.27	-0.07	0.25	-0.12	0.61	0.43	-0.01	0.16	-0.08	1.00	0.23
Tobacco Use Estimate	0.09	-0.49	-0.02	0.12	0.19	-0.01	0.38	0.30	-0.17	0.25	0.01	0.23	1.00

	Density (P/Km2)	Birth Rate	CPI	CPI Change (%)	GDP	Gross primary education enrollment (%)	Gross tertiary education enrollment (%)	Life expectancy	Total tax rate	Unemployment rate	Urban_population	Alcohol Total Per Capita Consumption	Tobacco Use Estimate
Density (P/Km2)	1.00	-0.14	-0.07	-0.07	-0.01	-0.02	0.11	0.16	-0.09	-0.11	-0.01	-0.14	-0.00
Birth Rate	-0.14	1.00	0.32	0.11	-0.19	0.08	-0.77	-0.88	0.16	-0.10	-0.09	-0.55	-0.47
CPI	-0.07	0.32	1.00	0.56	-0.07	0.28	-0.26	-0.39	0.01	0.06	0.06	-0.25	-0.10
CPI Change (%)	-0.07	0.11	0.56	1.00	-0.02	0.09	-0.02	-0.17	0.12	0.06	0.07	-0.05	0.03
GDP	-0.01	-0.19	-0.07	-0.02	1.00	-0.03	0.14	0.17	0.12	-0.06	0.89	0.06	0.07
Gross primary education enrollment (%)	-0.02	0.08	0.28	0.09	-0.03	1.00	-0.15	-0.10	0.02	-0.04	0.01	-0.10	0.01
Gross tertiary education enrollment (%)	0.11	-0.77	-0.26	-0.02	0.14	-0.15	1.00	0.77	-0.01	0.11	0.03	0.60	0.35
Life expectancy	0.16	-0.88	-0.39	-0.17	0.17	-0.10	0.77	1.00	-0.10	-0.02	0.03	0.41	0.32
Total tax rate	-0.09	0.16	0.01	0.12	0.12	0.02	-0.01	-0.10	1.00	-0.04	0.13	-0.10	-0.11
Unemployment rate	-0.11	-0.10	0.06	0.06	-0.06	-0.04	0.11	-0.02	-0.04	1.00	-0.04	0.14	0.17
Urban_population	-0.01	-0.09	0.06	0.07	0.89	0.01	0.03	0.03	0.13	-0.04	1.00	-0.04	0.06
Alcohol Total Per Capita Consumption	-0.14	-0.55	-0.25	-0.05	0.06	-0.10	0.60	0.41	-0.10	0.14	-0.04	1.00	0.25
Tobacco Use Estimate	-0.00	-0.47	-0.10	0.03	0.07	0.01	0.35	0.32	-0.11	0.17	0.06	0.25	1.00

Images 6,7: Spearman and Pearson correlation matrix heatmaps of all numerical features

Both heatmaps show that birth rate - life expectancy, tertiary education - life expectancy, GDP - urban population are correlated which is to be expected due to their relation to a country’s development level which lends more credibility to the datasets used.

Alcohol consumption is seemingly correlated with tertiary education, life expectancy and birth rate; thus, these correlations will be explored further during hypothesis testing. Alcohol consumption is also not that correlated with urban population, unemployment and tax rate implying that it is more dependent on long term social conditions rather than short term economic ones.

Tobacco usage is also correlated with birth rate, tertiary education and life expectancy although with weaker coefficients as compared to alcohol. In fact, tobacco usage has weaker correlations in general with almost every feature as compared to alcohol which is likely due to the stricter policies regarding its usage.

These heatmaps helped formulate the hypothesis tests that were conducted and will be discussed in the next section.

3.3 Hypothesis Tests

3.3.1 Test 1

The first test was performed between Gross Tertiary Education Enrollment and Alcohol Total Per Capita Consumption with the hypothesis:

Null Hypothesis (H_0): Gross Tertiary Education Enrollment and Alcohol Total Per Capita Consumption are not correlated.

Alternative Hypothesis (H_1): Gross Tertiary Education Enrollment and Alcohol Total Per Capita Consumption are correlated.

The test was conducted using Spearman’s Rank Correlation with the results giving an r-value of 0.6144 and p-value of 0.00000000000004070576016826558100420447533859. Thus, the null hypothesis was rejected, and the correlation was determined to be statically significant with $p < 0.05$ and positive and strong with $0.6 < |r| < 0.8$.

This correlation is likely due to education correlating with higher income and urban lifestyle along with drinking norms introduced with higher education

3.3.2 Test 2

The second test was performed between Gross Tertiary Education Enrollment and Tobacco Use Estimate with the hypothesis:

Null Hypothesis (H_0): Gross Tertiary Education Enrollment and Tobacco Use Estimate are not correlated.

Alternative Hypothesis (H_1): Gross Tertiary Education Enrollment and Tobacco Use Estimate are correlated.

The test was conducted using Spearman's Rank Correlation with the results giving an r-value of 0.3809 and p-value of 0.00001381036884722625871640966366271641163621. Thus, the null hypothesis was rejected, and the correlation was determined to be statistically significant with $p < 0.05$ and positive and weak with $0.2 < |r| < 0.4$.

This is likely due to a similar reason to the previous test with education correlating with higher income and urban lifestyle which in turn increases tobacco usage.

3.3.3 Test 3

The third test was performed between Unemployment Rate and Alcohol Total Per Capita Consumption with the hypothesis:

Null Hypothesis (H_0): Unemployment Rate and Alcohol Total Per Capita Consumption are not correlated.

Alternative Hypothesis (H_1): Unemployment Rate and Alcohol Total Per Capita Consumption are correlated.

The test was conducted using Spearman's Rank Correlation with the results giving an r-value of 0.1601 and p-value of 0.07700070031427859451600426154982415027916431. Thus, the null hypothesis could not be rejected with $p > 0.05$, and no statistically significant correlation was found.

As mentioned previously, this was likely due to unemployment rate not being enough to represent the stress levels of the population by itself and short-term economic stress not being likely to cause an effect on alcohol consumption.

3.3.4 Test 4

The fourth test was performed between Unemployment Rate and Tobacco Use Estimate with the hypothesis:

Null Hypothesis (H_0): Unemployment Rate and Tobacco Use Estimate are not correlated.

Alternative Hypothesis (H_1): Unemployment Rate and Tobacco Use Estimate are correlated.

The test was conducted using Spearman's Rank Correlation with the results giving an r-value of 0.2472 and p-value of 0.00583339327387159475857281165644963039085269. Thus, the null hypothesis was rejected, and the correlation was determined to be statistically significant with $p < 0.05$ and positive and weak with $0.2 < |r| < 0.4$.

This correlation shows that the stress caused by unemployment does affect tobacco usage, albeit weak.

3.3.5 Test 5

The fifth test was performed between Birth Rate and Alcohol Total Per Capita Consumption with the hypothesis:

Null Hypothesis (H_0): Birth Rate and Alcohol Total Per Capita Consumption are not correlated.

Alternative Hypothesis (H_1): Birth Rate and Alcohol Total Per Capita Consumption are correlated.

The test was conducted using Spearman's Rank Correlation with the results giving an r-value of -0.5935 and p-value of 0.00000000000046591517730011336083978090526297. Thus, the null hypothesis was rejected, and the correlation was determined to be statistically significant with $p < 0.05$ and negative and moderate with $0.4 < |r| < 0.6$.

This negative correlation is likely due to the correlation of lower income with higher birth rate and the cultures of countries with higher birth rates discouraging alcohol use.

3.3.6 Test 6

The sixth test was performed between Birth Rate and Tobacco Use Estimate with the hypothesis:

Null Hypothesis (H_0): Birth Rate and Tobacco Use Estimate are not correlated.

Alternative Hypothesis (H_1): Birth Rate and Tobacco Use Estimate are correlated.

The test was conducted using Spearman’s Rank Correlation with the results giving an r-value of -0.4909 and p-value of 0.00000000820741608105041937865484815697522669. Thus, the null hypothesis was rejected, and the correlation was determined to be statistically significant with $p < 0.05$ and negative and moderate with $0.4 < |r| < 0.6$.

The reason for this correlation most likely occurs from the same reasons with the previous test with discouragement of tobacco usage just like alcohol.

3.3.7 Test 7

The seventh test was performed between Alcohol Regulation and Alcohol Total Per Capita Consumption with the hypothesis:

Null Hypothesis (H_0): There is no difference in Alcohol Total per Capita Consumption between countries with and without regulation.

Alternative Hypothesis (H_1): Alcohol Total per Capita Consumption differs between countries with and without regulation.

The test was conducted using Mann-Whitney U with the results giving a p-value of 0.0534646548928092 7. Thus, the null hypothesis could not be rejected with $p > 0.05$, and no statistically significant difference was found in alcohol consumption between countries with and without alcohol regulations.

This is likely due to alcohol not being heavily regulated and the regulation factor by itself being insufficient to show any differences.

3.3.8 Test 8

The eighth test was performed between Tobacco Regulation and Tobacco Use Estimate with the hypothesis:

Null Hypothesis (H_0): There is no difference in Tobacco Use Estimate between countries with and without regulation.

Alternative Hypothesis (H_1): Tobacco Use Estimate differs between countries with and without regulation.

The test was conducted using Mann-Whitney U with the results giving a p-value of 0.004870622044438068. Thus, the null hypothesis was rejected with $p < 0.05$, and countries with tobacco regulations were determined to have statistically significantly higher tobacco use estimates than those without regulations.

This counter-intuitive result was discussed previously and is due to regulations being reactive measures instead of preventative measures.

3.4 Machine Learning

Several regression type models were trained on the dataset to find the best predictive model on alcohol consumption and tobacco usage based on the remaining features. The dataset was first prepared for training by replacing “Yes” and “No” values with 1’s and 0’s along with dropping the “Country” column. Then the targets were set as “Alcohol Total Per Capita Consumption” and “Tobacco Use Estimate” and the remaining features were set as predictors. All models trained used 10-fold cross validation due to the small size of the dataset.

3.4.1 Linear, Ridge and Lasso Models

Linear, Ridge and Lasso models were trained first with Lasso and Ridge models using standard regularization parameters and their results were as follows:

Model	Target	RMSE	R^2
Linear	Alcohol	3.4334	0.2626
Ridge	Alcohol	3.3766	0.2868
Lasso	Alcohol	3.3944	0.2793
Linear	Tobacco	8.9059	0.1273
Ridge	Tobacco	8.8825	0.1319
Lasso	Tobacco	8.8711	0.1341

The linear regression model only provides a baseline for both targets, especially considering the performance of the model. The model had high error values on both alcohol and especially tobacco with low R^2 and high RMSE values showing that there are weak linear relationships between the predictors and targets and such linear relationships are insufficient to predict the targets.

The ridge regression model showed a mild improvement over linear regression indicating a mild multicollinearity among the predictors. But due to the model still being based on linear relationships, it had high error values.

Lasso regression performed similarly to the ridge regression model with a slightly better performance on tobacco usage implying that there might be some redundant predictors.

3.4.2 K-Nearest Neighbors (kNN) Model

kNN models were trained on both predictors for k between 1 and 20 with the following RMSE over different k values [8].

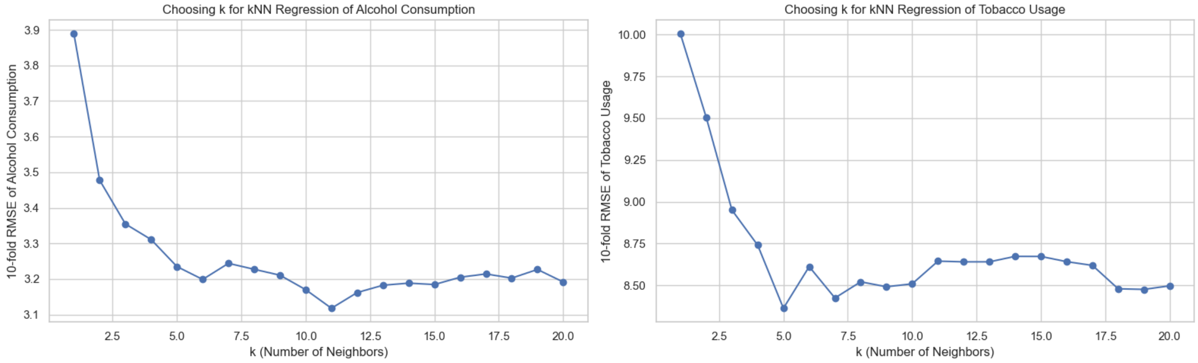


Image 8: Graphs showing RMSE values for different k’s for Alcohol Consumption and Tobacco Usage

The best k for alcohol consumption was determined to be 11 with a 3.1179 RMSE and 0.3963 R^2 value. For tobacco, best k was 5 with 8.3632 RMSE and 0.2326 R^2 values.

The model performed better than the linear regression models for both targets which implies the existence of a nonlinear structure. The difference in k values between the targets shows that alcohol consumption has a smoother trend which resulted in a larger best k compared to tobacco usage having more local variability leading to a smaller best k.

3.4.3 Decision Tree Model

Decision tree models were trained on both predictors with maximum depth between 1 and 20 the following RMSE over different maximum depths [9].

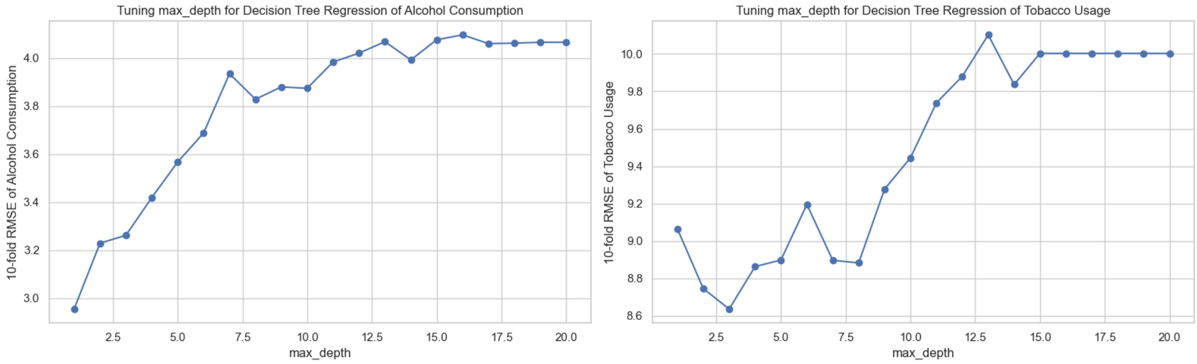


Image 9: Graphs showing RMSE values for different maximum depths for Alcohol Consumption and Tobacco Usage

The best maximum depth for alcohol consumption was determined to be 1 with 2.9548 RMSE and 0.4543 R^2 values. For tobacco usage, best maximum depth was 3 with 8.6351 RMSE and 0.1772 R^2 values.

The results show that very shallow trees performed best with best maximum depth 1 for alcohol and 3 for tobacco. This seemingly reveals that there are no complex or hierarchical interactions between the features and their relationships are mostly monotonic or linear. The models revealed that the underlying structures are simple but single trees are too robust to properly explain them. The performance of the tree is also limited due to high variance.

3.4.4 Random Forest Model

Random forest models were trained on both predictors hoping to improve over single decision trees and kNN with the hyperparameters max_depth = [3,5,10,15], n_estimators = [10,50,100]. For alcohol the best hyperparameters and results were:

- max_depth = 15
- n_estimators = 100
- RMSE = 2.8411
- R^2 = 0.4965

For tobacco the best hyperparameters and results were:

- max_depth = 15
- n_estimators = 100
- RMSE = 7.8373
- R^2 = 0.3204

The model showed a significant improvement over previous models, likely due to it reducing the variance of individual trees, its ability to capture weak nonlinear relationships and it being robust to noise and feature interactions. The best hyperparameters being identical for both targets indicate that both targets benefit from relatively deep trees when variance is controlled and that the findings about the underlying structure from the decision tree model may not be so accurate.

3.4.5 XGBoost Model

The final model trained was the XGBoost model with hyperparameters max_depth = [2,3,4], n_estimators = [50,100,200] and learning_rate = [0.03,0.05,0.1]. The results and the best hyperparameters for alcohol were:

- learning_rate = 0.03
- max_depth = 4
- n_estimators = 200
- RMSE = 2.7246
- R^2 = 0.5356

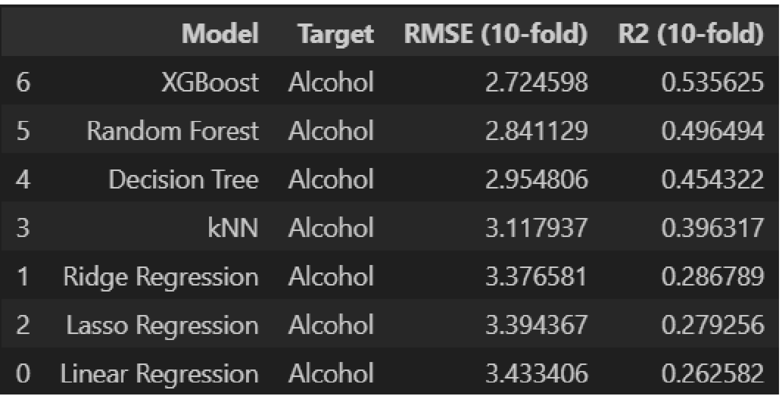
The results and hyperparameters for tobacco were:

- learning_rate = 0.05
- max_depth = 4
- n_estimators = 200
- RMSE = 7.6254
- R^2 = 0.3602

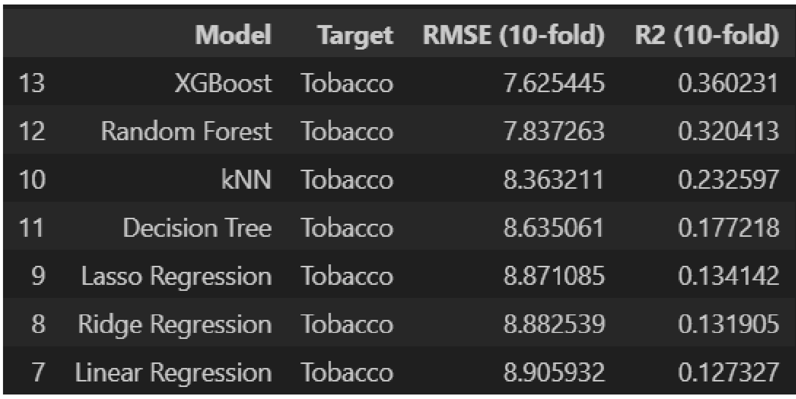
The XGBoost models were the best performing models even outperforming the random forest models, likely due to them better handling the weak and structured nonlinearities.

4. Findings

As mentioned above, the best performing model on both Alcohol Consumption and Tobacco Usage was XGBoost. The comprehensive tables showing each model's performance predicting the targets sorted by RMSE are below [10,11]:



	Model	Target	RMSE (10-fold)	R2 (10-fold)
6	XGBoost	Alcohol	2.724598	0.535625
5	Random Forest	Alcohol	2.841129	0.496494
4	Decision Tree	Alcohol	2.954806	0.454322
3	kNN	Alcohol	3.117937	0.396317
1	Ridge Regression	Alcohol	3.376581	0.286789
2	Lasso Regression	Alcohol	3.394367	0.279256
0	Linear Regression	Alcohol	3.433406	0.262582



	Model	Target	RMSE (10-fold)	R2 (10-fold)
13	XGBoost	Tobacco	7.625445	0.360231
12	Random Forest	Tobacco	7.837263	0.320413
10	kNN	Tobacco	8.363211	0.232597
11	Decision Tree	Tobacco	8.635061	0.177218
9	Lasso Regression	Tobacco	8.871085	0.134142
8	Ridge Regression	Tobacco	8.882539	0.131905
7	Linear Regression	Tobacco	8.905932	0.127327

Images 10,11: Tables showing performance of different models on predicting Alcohol and Tobacco Consumption sorted by RMSE

The XGBoost models can explain around 54% of the variance in Alcohol Consumption and around 36% variance in Tobacco Usage based on the R^2 values. These values are to be expected as in this domain such values are common due to several factors such as the noise found in real world data and even large models trained by WHO rarely exceed 0.5 R^2.

The importances of features for the XGBoost models are shown below [12]:

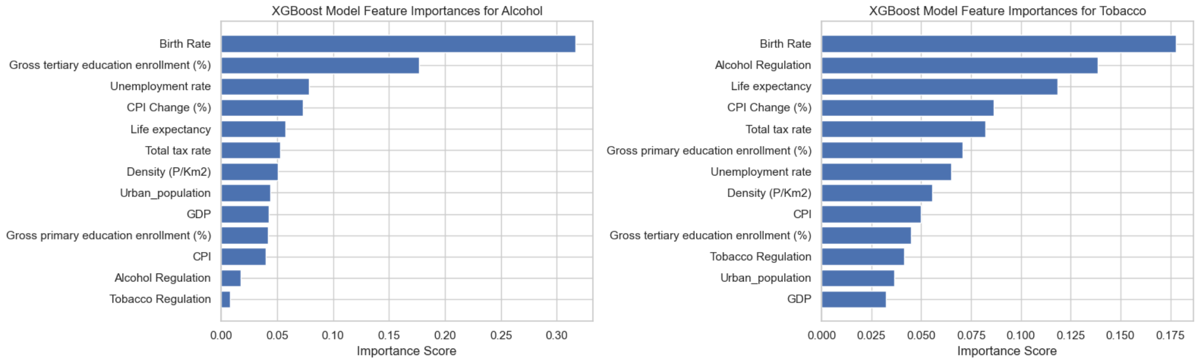


Image 12: Graphs showing feature importance scores for the XGBoost models of Alcohol Consumption and Tobacco Usage

The feature importance scores for predicting alcohol consumption reveal that the biggest predictor is birth rate. Birth rate strongly affecting alcohol consumption is likely linked to the demographic structure of the community with lower birth rates indicating aging populations that can consume alcohol and generally more developed societies.

Another important factor, gross tertiary education enrollment is likely due to it being an indicator for developed countries with higher incomes and the drinking norms associated with tertiary education.

Unemployment rate, CPI change, life expectancy and total tax rate have comparably lower importance, but they suggest that alcohol consumption is sensitive to economic stress and stability along with overall health and longevity.

Regulations have minimal importance on alcohol consumption compared to socio-economic factors implying that they cannot explain cross-country alcohol consumption differences by themselves and consumption is more linked to culture and demographics rather than regulations.

The tobacco graph shows that once again the most important feature in predicting tobacco usage is birth rate, likely due to the same reasons for alcohol. Its importance is less dominant compared to alcohol which indicates that the demographic effect on tobacco usage might be weaker.

A surprising factor with high importance is alcohol regulation which is likely due to policy spill over which means that countries that have alcohol regulations are most likely to have stricter public health norms or better enforcement cultures which lead to tobacco usage getting affected as well.

Life expectancy is another important predictor, likely due to the relationship between smoking and population health.

CPI change, total tax rate, gross primary education enrollment and unemployment rate have medium importance over the predictions indicating tobacco usage is more sensitive to economic pressure like alcohol along with the basic education level of the country.

Tobacco regulations themselves have minimal influence over the prediction which is most likely caused by almost all the countries featured in the dataset having some form of regulation in place which then reduces the predictive power of the feature.

5. Limitations and Future Work

The biggest limitation of the analysis is that the dataset used is based on a single year, 2023, leading to the absence of temporal data and their effects and it has a low size with only 123 entries. This may lead to the models underperforming or finding non-existent relationships due to increased sensitivity to noise. This was mitigated as much as possible through the usage of 10-fold cross validation, but the usage of larger datasets is more certain as the performance of the models may have been affected depending on the data split.

Another limitation is that there may be more relevant features contributing to the consumption of alcohol and usage of tobacco that were not present in this study such as the happiness level of a country. Possibly the quality of the features could also be increased as the predictors were aggregated, country-level data. Such data may mask sub-national heterogeneity and may miss crucial patterns.

Feature works may address these limitations and build larger and more accurate models that can predict what contributes to the usage of these substances.