The Future of Alcohol Consumption

Kyle Mulkins, William Caras, and Conor Desmond

Mathematics & Computer Science, Marist College

Data 477: Data Science Capstone

Professor Kienzle

May 3, 2024

Data Capstone Final Project
Kyle Mulkins, William Caras, & Conor Desmond

Our Abstract

In the midst of a mounting public health concern, our interdisciplinary team of data science majors has embarked on a mission to elucidate the intricate pathways linking alcohol consumption and its repercussions on human life in the United States. Our project employs predictive analytics and visualization techniques to address alcohol-related mortality comprehensively. Our primary objective is to leverage advanced machine learning methodologies to accurately forecast alcohol-related deaths, transcending the limitations of traditional statistical approaches.

At the heart of our endeavor is the thorough analysis of extensive datasets. We meticulously filter out extraneous information to present a lucid depiction of alcohol's multifaceted effects. By harnessing the power of predictive analytics, we aspire to anticipate alcohol-related risks and empower health organizations to intervene effectively. Moreover, our research endeavors to equip students with insights that can potentially shape their future decisions and contribute to a healthier society. Through this approach, we aim to shed light on the complex interplay between alcohol consumption and mortality, ultimately contributing to the development of targeted. Interventions and strategies for mitigating the adverse impacts of alcohol misuse on public health.

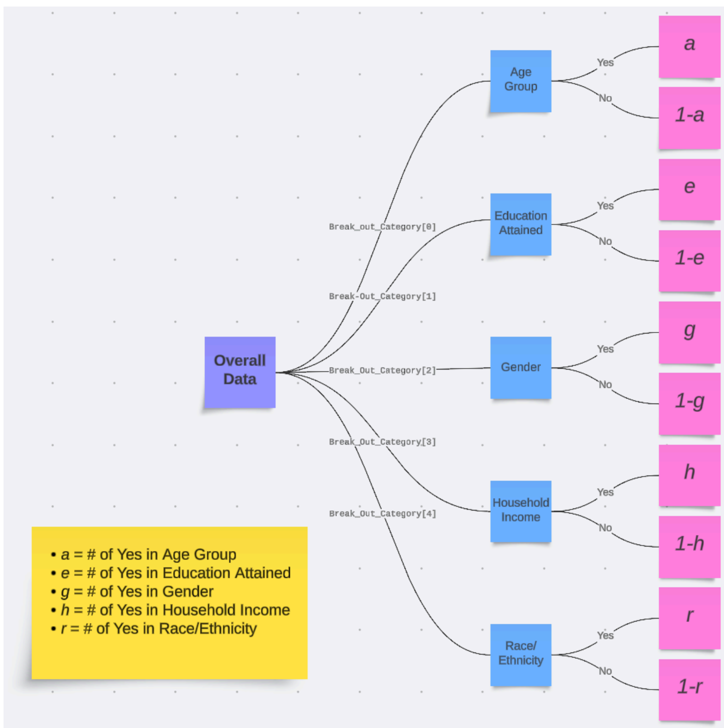
Data Capstone Final Project
Kyle Mulkins, William Caras, & Conor Desmond

Value

Our project consists of analyzing the consumption of alcohol on many different demographics, including race, gender, age, geographical location, and more. Alcohol has been a major component in the lives of Americans. College students find themselves out at bars weekly, (even though they cannot admit it). Middle-aged people find alcohol as a great pastime when they're feeling low. Finally, the older generation drinks alcohol too, mainly because it was something they've done for most of their lives. Race and ethnicity also should have a major effect, as different types of alcohol are characteristically compared to different nationalities (Vodka to Russians, Tequila to Mexicans, etc…). Different states can have varying rates of alcohol consumption per capita too! Don't forget about household income as well. The amount of parameters for this topic is mind-blowing, and knowing how alcohol consumption differs for each person can show who is the most likely to drink, or binge drink, or even get alcohol poisoning. This analysis is meant to show these different parameters, and how each of them can predict the future of alcohol consumption, so hopefully we can prevent a disaster before one takes the wheel.

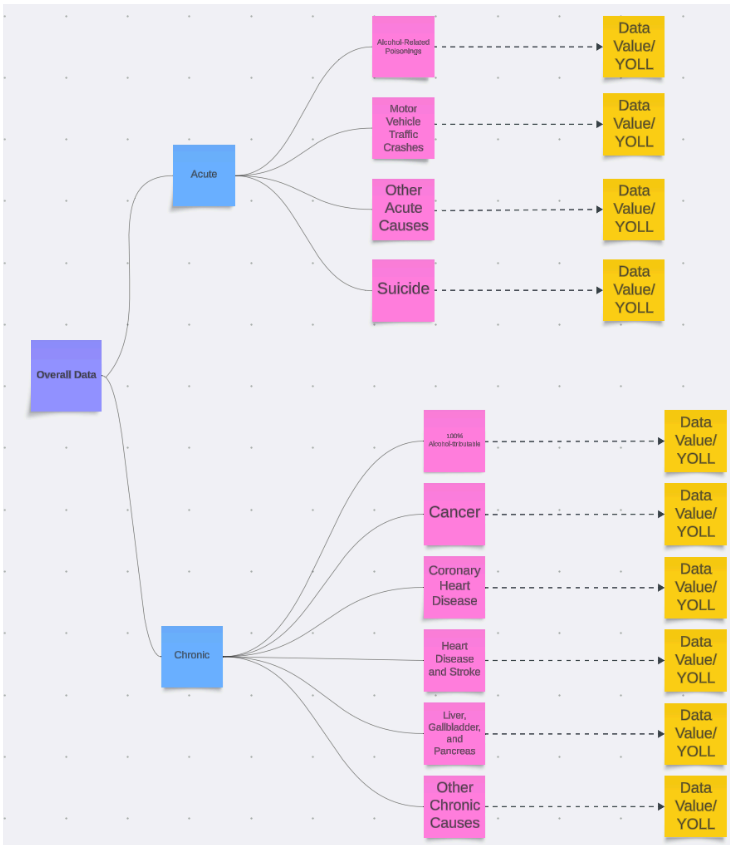
User Experience

If someone would like to use our models to make further predictions or make analysis on other datasets, here is how a user may utilize our models. Firstly, We include a wide variety of models that predict the values of many different aspects of the 3 datasets we included. For example, suppose a user wants to predict the future responses to the BRFSS survey dataset based on the data of previous years. What's interesting about the models within this dataset is the inclusion of break-out categories. The break-out categories split up the data of a given year into different categories like race, gender, household income, etc. This system of subcategories of the original data can allow us to make more specific models. So the user can categorize the data based on a certain break-out category and make further analysis on each group, creating more specific models. One of the more important models with this dataset is the Naive Bayes model, where a user may do what is described above:



Notice that by understanding the format of the break-out categories, we can create 5 separate models looking into specifically the age, education, gender, income, and race/ethnicity of the participants of the survey. All of this leads to results specific to the given category.

The user can also experiment similar to above with the ARDI dataset as well. While the BRFSS had more categories to analyze, the ARDI dataset has a well-defined structure that makes it easier to grab age ranges, or even causes of death.



Notice that the first column after the overall data (the blue boxes) represent the condition type variable, which can be either acute or chronic. Further into the next column (the pink boxes) represent the category of the death, which includes 10 different types of disease, injuries and ailments. There are further attributes to look into, like the specific cause of death and the age range for each, which is not shown in the flowchart. All of these would lead to either the death toll of that record (data value) or the Years of Potential Lost Life for that record (YOLL). A user may use these features in order to get predictive models on specific elements of a person's death, and see which of these have the biggest factors in causing death.

Overall, a user can use our project to look into the data of alcohol and how it has impacted those in the United States. With different datasets, a user can also do the same thing with other countries or even a global look into alcohol's effects. For an admin though, they may consider creating further models that explain different aspects of other countries. For example, an admin can consider Ireland’s drinking (which is younger and has a lot of beer) and how that affects their death toll, their car accidents, or even track the amount of alcohol that will be consumed in the future depending on the day. Also, the admin may look into further studies on how alcohol affects daily life, meaning when the alcohol consumed is likely business-related (like talking to clients or work parties) or for personal gain (like drinking with friends, going to bars, or drinking solo). This could further explain the demographic more likely to cause deaths related to alcohol or answer yes to the BRFSS questions. These ideas are some ideas an admin could get by using what we have in our models, and how they can find further results than us.

Process

Dataset Description and Variables:

The first dataset we used was from the Alcohol-Related Disease Impact Application. The Alcohol-Related Disease Impact (ARDI) Application offers national and state-level estimates of the health impacts associated with alcohol consumption, specifically focusing on deaths and years of potential life lost (YPLL). These estimates are derived using alcohol-attributable fractions and cover a range of 58 acute and chronic causes. The data is reported by age and sex from 2015 to 2019. The primary focus is on quantifying the total number of deaths that can be attributed to alcohol consumption from various factors, including alcohol poisoning, pneumonia, injuries, or even cancer.

- `YearEnd` - The year in which the study ended. The study began in 2015 and ended in 2019.
- `LocationAbbr` - The US State 2 letter abbreviation of the current record
- `LocationDesc` - The full US State name of the current record
- `DataSource` - The source from which the data was collected. All records are from ARDI
- `ConditionType` - The general type of condition that led to the death of a given record. Either Acute or Chronic
- `Category` - The general category of the death of a given record
- `Cause_of_Death` - The accepted cause of death of a given record
- `Data_Value_Unit` - The units of the data value we are measuring. Since YOLL is added, this column is useless.
- `Data_Value_Type` - The measurement value in which the number of deaths are being tracked
- `Data_Value_Alt` - The actual data value we are measuring (Has many nulls, which is why YOLL was added)
- `Effect` - The effect that alcohol had on the people with the given condition in the record. Either Harmful or Beneficial.
- `ConsumptionPattern` - The measure of how much alcohol is being consumed. Either Any Alcohol Use or Excessive Alcohol Use.
- `Sex` - The sex of the people in the given record. Either Male, Female, or Overall
- `AgeCategory` - A generalized category that explains if the people in the record were under 21 or not.

- `AgeGroup` - An age range given to the people who died from the given condition. Can be a range of years, or Overall.
- `YOLL` - The measurement of Years of Potential Lost Life, which estimates the amount of years of lost life caused by the people who have died in this record.

Our second dataset is from the Behavioral Risk Factors Surveillance System. The Behavioral Risk Factors Surveillance System (BRFSS) provides a table of alcohol consumption data from 2011 to the present. The data is collected through landline and cell phone surveys, representing a continuous, state-based surveillance system. BRFSS focuses on gathering information about modifiable risk factors for chronic diseases and other leading causes of death. The data is updated annually as new information becomes available. For detailed information on the sampling methodology and quality assurance processes, additional resources are available on the BRFSS website.

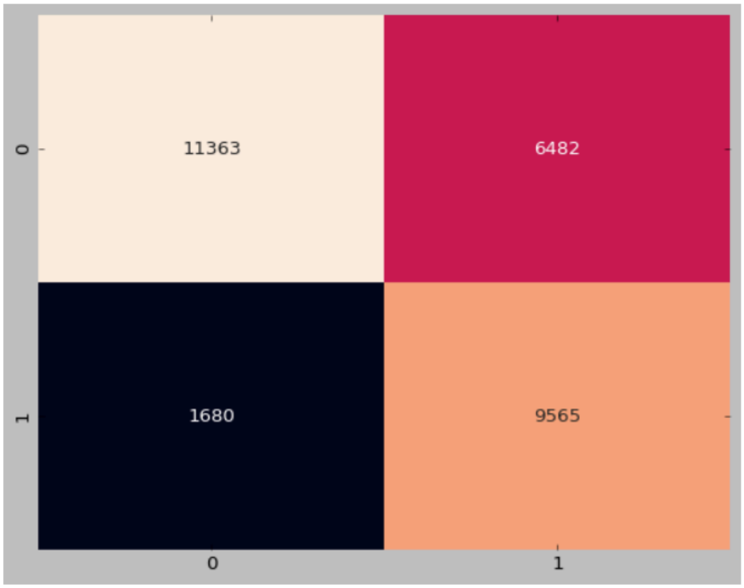
- `Year` - The year in which the survey question was asked. From 2011-2022
- `Locationabbr` - The US State 2 letter abbreviation of the current record
- `Locationdesc` - The full US State name of the current record
- `Class` - The class of the question being asked. All Alcohol Consumption
- `Topic` - The topic of the question being asked. Either Alcohol Consumption, Binge Drinking, or Heavy Drinking
- `Question` - The question being asked in the survey. One of 3 different questions.
- `Response` - The response to the question given. Either Yes or No
- `Break_Out` - The variable measured by the break out category. Different values exist for different break out categories.

- `Break_Out_Category` - The category being measured from this question. Can be Age Group, Education Attained, Gender, Household Income, Race/Ethnicity, or Overall
- `Sample_Size` - The amount of people who fit the given demographic of the break out category.
- `Data_value` - The percentage of people who said Yes or No based on the demographic of the break out category.
- `Confidence_limit_Low` - The lower limit of the confidence interval for the approximate data value.
- `Confidence_limit_High` - The upper limit of the confidence interval for the approximate data value.
- `Display_order` - Explains the percentage of the entire break out category that this data value measures.
- `Data_value_unit` - The units of the measurement values. Always percentages.

Using this dataset, Location abbreviation, Class, Topic, Data value, Confidence limit Low, Confidence limit High, Display order, and Data value unit were dropped. Dummy values, (0 and 1), were used to replace the values in the variables that could be binary. This was done so these variables could be used in Naive Bayes models. If there were more than 2 different values in the variable then new variables were made of each of the column's values and a 0 was put for each of these new variables if the value the variable represented was not in the row and a 1 was put if the value was in the row. A Response Yes variable was added to be the target. This variable has a 0 if the response was No and a 1 if the response was Yes. The Response No variable was dropped since that should not be an input since it is the opposite of the Response Yes variable. Bernoulli Naive Bayes classifier was made to predict the target. It had 59.6% precision, 72% accuracy, and 74.4% AUC. The precision means only

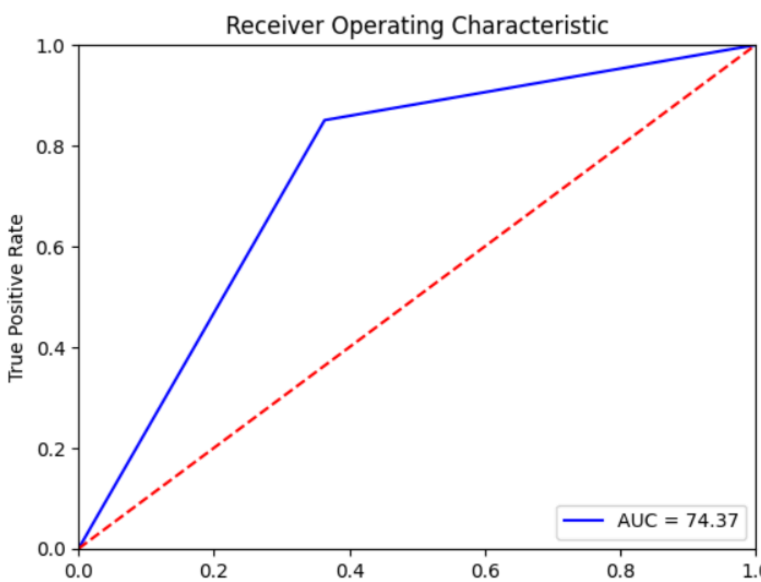
59.6% of the values predicted to be Yes were actually Yes. The AUC of 74.4 means there is a 74.4% chance our model can correctly identify a Yes value and a No value if one of each were pulled and given to the model. The recall is 85.1%, which means about 85.1% of Yes values were predicted to be Yes by the model. The specificity is 63.7%, which means about 63.7% of No values were predicted to be No by the model. The f1 score is “the harmonic mean of precision and recall” (“F1 Score” para. 2). The f1 score is 70.1% which is pretty good. A categorical naive bayes classifier without year and sample size was made. Sample size was not included since the model would not work with sample size and year was not used; it would likely not be important, like in Bernoulli Naive Bayes. This classifier had 50.3% precision, 62% accuracy, and 68.9% AUC. The precision means only 50.3% of the values predicted to be Yes were actually Yes. The AUC of 68.9 means there is a 68.9% chance our model can correctly identify a Yes value and a No value if one of each were pulled and given to the model. The recall is 100%, which means 100% of Yes values were predicted to be Yes by the model. The specificity is 37.8%, which means about 37.8% of No values were predicted to be No by the model. The f1 score is 67% which is pretty good. These models could be used with pretty good accuracy to predict if a lot of people will drink alcohol in excess.

<u>Bernoulli Naive Bayes Results:</u>

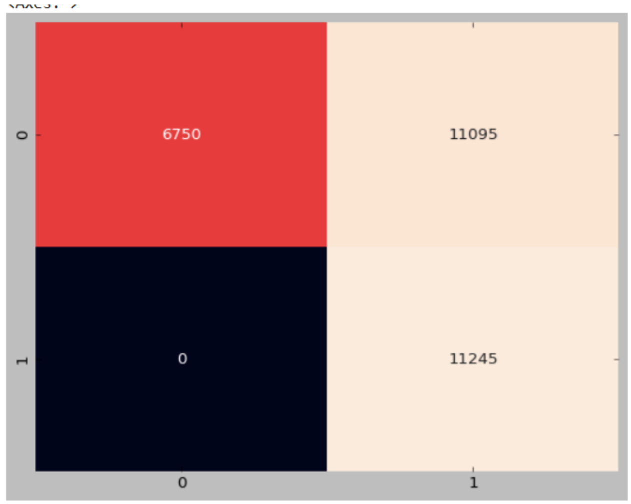


precision: 59.6%
recall: 85.1%
specificity: 63.7%
roc: 74.4%
f1: 70.1%

	precision	recall	f1-score	support
0	0.87	0.64	0.74	17845
1	0.60	0.85	0.70	11245
accuracy			0.72	29090
macro avg	0.73	0.74	0.72	29090
weighted avg	0.76	0.72	0.72	29090

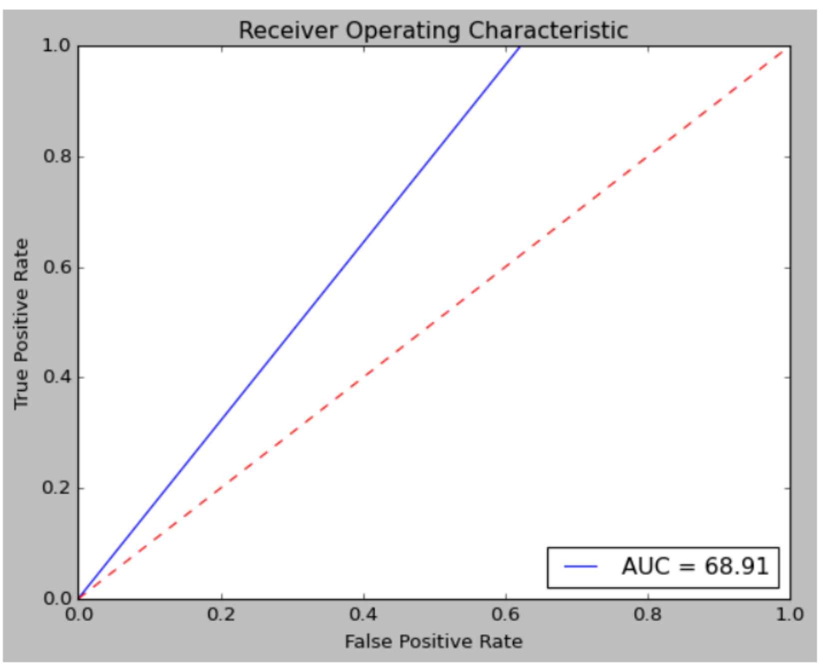


<u>Categorical Naive Bayes Results:</u>




```
precision: 50.3%
recall: 100.0%
specificity: 37.8%
roc: 68.9%
f1: 67.0%
```

	precision	recall	f1-score	support
0	1.00	0.38	0.55	17845
1	0.50	1.00	0.67	11245
accuracy			0.62	29090
macro avg	0.75	0.69	0.61	29090
weighted avg	0.81	0.62	0.60	29090



Our third and final dataset is a general dataset on Alcohol Consumption over the years from 1977-2018 in each US state. This dataset is created by Jacob Kaplan from the University of

Pennsylvania. This data set contains the per capita consumption of ethanol (in gallons) for each state in the US for the years 1977-2018, later simplified to 1984-2018 due to the change of the age of drinking. This dataset includes the total ethanol consumed and the consumption by three categories: beer, wine, and spirits. The same categories are given to data tracking the number of drinks consumed.

- `state` - The US state that is being measured.
- `year` - The year that is being measured.
- `ethanol_beer_gallons_per_capita` - The amount of ethanol from beer being consumed in a given state and year.
- `ethanol_wine_gallons_per_capita` - The amount of ethanol from wine being consumed in a given state and year.
- `ethanol_spirit_gallons_per_capita` - The amount of ethanol from spirits being consumed in a given state and year.
- `ethanol_all_drinks_gallons_per_capita` - The amount of ethanol from all drinks being consumed in a given state and year.
- `number_of_beers` - The number of beers being consumed in a given state and year.
- `number_of_glasses_wine` - The number of glasses of wine being consumed in a given state and year.
- `number_of_shots_liquor` - The number of shots of liquor being consumed in a given state and year.
- `number_of_drinks_total` - The total number of drinks being consumed in a given state and year.

A variable for total drinks consumed in the next year by each state was added to the Alcohol Consumption over the years data set to be the target. Rows were dropped where the state variable was regions that were not states in order to make a model that predicted total drinks consumed in the next year by state. Rows were dropped where the year was 2018 because we had no data for drinks consumed in 2019. For the input variables data frame (df_alc2), the state variable and the target were dropped. State was a text variable so that's why it was dropped. The input data was scaled using a standard scaler. After that, Pearson's correlation was used to get the inputs that had at least a 0 correlation with the target. 0 was used since we only had 8 inputs and we wanted to maximize our number of input variables. The inputs and target were split from the non scaled data, and the input data was scaled. PCA and LLE were run on the input data, but these sets did not provide the best models, so they were not used. Total drinks consumed in the next year for a state was predicted using many different models: Ridge regression, Lasso regression, stochastic gradient descent, gradient boosting regressor, decision tree regressor, support vector machine with a rbf kernel, support vector machine with a poly kernel, support vector machine with a linear kernel, linear support vector regression, elastic regression, bagging regressor, and keras neural network. These algorithms were chosen since we had functions for these already and they could help make the most accurate prediction. Some of the best models were taken based on their average normalized score for the test set and the difference in average normalized score between the training set and test set and tried to make a blender. The average normalized score is the normalized MAE and normalized RMSE averaged together. Eventually, we got a blender with 4 models that gave me the best average normalized score and difference in scores out of all the blenders. This blender has a 20% average normalized score for the test set, .5% difference in scores, and uses the non PCA or LLE data. Lasso regression, ridge regression, stochastic gradient descent, and elastic regression made up the blender and all used the non

PCA or LLE data. The lasso regression has a 19.9% average normalized score for the test set and .27% difference in scores. The ridge regression has a 19.9% average normalized score for the test set and .28% difference in scores. The stochastic gradient descent has a 19.9% average normalized score for the test set and .295% difference in scores. The elastic regression has a 19.9% average normalized score for the test set and .302% difference in scores. So, the scores are high, but understandable for only having 8 input variables and trying to predict consumption for any state, which is difficult. This shows us the importance of having more input variables. It seems like the best model is the lasso regression since it has the lowest difference in scores and best average normalized score. The libraries used are sklearn, pandas, pickle to save and load the SGD model, and numpy for random samples. The right picture is Lasso results, Ridge results, and then SGD results in that order respectively.

```
Blender Train Scores then Test Scores:
Avg. Normalized Score:19.5%

Avg. Normalized Score:20.0%
```

Elastic results below:

```
Avg. Normalized Score:19.6%
Test predictions:
Avg. Normalized Score:19.9%
Difference of avg scores:0.302%
```

```
Avg. Normalized Score:19.6%
Test predictions:
Avg. Normalized Score:19.9%
Difference of avg scores:0.269%
```

```
Avg. Normalized Score:19.6%
Test predictions:
Avg. Normalized Score:19.9%
Difference of avg scores:0.279%
```

```
Avg. Normalized Score:19.6%
Test predictions:
Avg. Normalized Score:19.9%
Difference of avg scores:0.295%
```

Lasso Regression: Lasso regression, L1 regularization, is a statistical method that improves model accuracy by performing variable selection and regularization. It shrinks the coefficients of less important features to zero, thus simplifying the model and preventing overfitting. Initially introduced in geophysics, lasso regression has been widely adopted in various fields for its ability to maintain model interpretability while improving prediction accuracy.

Ridge Regression: Ridge regression, also known as L2 regularization, is a technique used to address multicollinearity in linear regression models by adding a penalty equal to the square of the magnitude of coefficients. This penalty term shrinks the coefficients and helps reduce model complexity, reducing overfitting on training data. It is beneficial when there are more predictors than observations or when several predictors correlate. Despite the shrinkage of the coefficients, ridge regression provides unbiased, stable, and reliable estimates, making it a valuable tool in predictive modeling.

Elastic (Elastic Net) Regression: is a regularization technique that combines the strengths of Lasso (L1) and Ridge (L2) regression methods. It is beneficial for models with numerous features, effectively reducing overfitting by penalizing the size of all coefficients and encouraging a sparse model representation. This method helps in feature selection and shrinking coefficients, improving model prediction accuracy and interpretability. Elastic Net is versatile and can be applied to various predictive modeling tasks, especially when dealing with high-dimensional datasets where the number of features exceeds the number of observations.

Stochastic Gradient Descent: The gradient is “a vector pointing in the general direction of the function’s steepest rise” (Stochastic Gradient Descent para. 2) at a point. So, the gradient is a vector that is the steepest slope of a function. “Gradient Descent is an iterative optimization process that searches for an objective function’s optimum value” (Min or Max) (Stochastic

Gradient Descent para. 1). So, the function loops through the data to search for a function's min or max. It changes the model's parameters to reduce the model's cost function. But, the function may need to iterate many times to find the max. or min. Stochastic gradient descent addresses this computational inefficiency. In Stochastic Gradient Descent, "instead of using the entire dataset for each iteration, only a single random training example (or a small batch) is selected to calculate the gradient and update the model parameters" (Stochastic Gradient Descent para. 5) This greatly reduces the cost for each iteration. This random selection introduces randomness into the optimization process, which is why the term "stochastic" is in "Stochastic Gradient Descent". Stochastic means randomly determined (Stochastic Gradient Descent para. 5).

Bernoulli Naive Bayes: It uses the Bayes Theorem to classify any event "based on the events that have already occurred" (Bernoulli Naive Bayes para. 2). It assumes that "all the events are independent" (Bernoulli Naive Bayes para. 2) and it is for the classification of binary features, and is used when one is given an event (Bernoulli Naive Bayes para. 5). The term naive comes from the naive assumption of the features not affecting each other, this is rarely true (Naive Bayes Classifiers para. 6).

Categorical Naive Bayes: "a collection of classification algorithms based on Bayes' Theorem" (Naive Bayes Classifiers para. 2). It is "a family of algorithms where all of them share a common principle:" (Naive Bayes Classifiers para. 2) all the pairs of features being classified are independent of each other. It "predicts the probability of an instance belonging to a class with a given set of feature values" (Naive Bayes Classifiers para. 5). So, it answers the question: given some values, what is the probability that this value belongs to a certain class? In the real world, the features rarely affect each other.

Future Directions

Expansion of Data Sources: Branching out into international comparisons of alcohol consumption. Data from other countries will be incorporated to compare mortality rates and identify global trends. This comparative analysis provides valuable insights into the effectiveness of various policies and interventions. For example, the drinking age affects consumption throughout the years. Is it better to be at the drinking age of 18 or 21, and why?

Integration of Socioeconomic Factors: Data Augment the analysis with socioeconomic indicators such as income, education level, and employment status to explore the relationships with alcohol-related mortality. Along with location, it identifies geographic hotspots that could guide allocating resources and interventions to areas with the greatest need. Understanding these associations could inform more equitable policies and future lives.

Advanced Visualization Techniques: Including visual aspects such as interactive dashboards. Develop interactive visualization dashboards to allow stakeholders, policymakers, and the general public to dynamically explore and interact with the data.

These dashboards could facilitate data-driven decision-making and increase morality issues. For example, these could highlight periods of significant change or involvement effectiveness.

By pursuing these future directions, our project can continue to advance our understanding of the complex interplay between alcohol consumption and mortality, ultimately contributing to more effective involvement and strategies for improving public health outcomes.

Ethical/Privacy Implications

Informed Consent and Data Privacy: Ensuring that individuals providing data for analysis are fully informed about the research's purpose, risks, and benefits. Explicit consent from participants is essential, particularly when using sensitive health-related data. Along with data privacy and implementing privacy measures to safeguard the confidentiality and security of individuals' personal information. Including anonymization techniques, encryption protocols, and access controls to prevent unauthorized disclosure or misuse of data.

Potential Harm and Stigmatization: Acknowledging the potential for harm associated with the analysis of alcohol-related mortality data. Care must be taken to avoid inadvertently exacerbating stigma or discrimination against individuals struggling with alcohol misuse or related health issues. Mitigating the risk of stigmatization by presenting findings in a sensitive and nonjudgmental manner. Avoiding the use of language or visualizations that perpetuate stereotypes or reinforce negative attitudes toward individuals with alcohol-related disorders.

Fairness and Bias Mitigation: Ensuring that the analysis and interpretation of data are conducted fairly and equitably, free from bias or prejudice. This includes addressing potential biases in the data collection process and mitigating algorithmic biases in predictive models. Implementing techniques such as fairness-aware machine learning algorithms and bias audits to identify and reduce biases in the data and models. Transparency about the limitations and potential biases of the analysis is essential for maintaining trust and credibility.

Responsible Use and Transparency: Committing to responsibly using data and research findings to minimize potential harm and maximize societal benefit. This includes adhering to ethical guidelines and regulations governing human subjects' data use in research. While providing clear and transparent communication about the research's purpose, methods, and implications to stakeholders, participants, and the broader community. Openness about data sources, analysis techniques, and potential limitations fosters trust and accountability.

Conclusion

Alcohol consumption is a major problem in the United States in recent history, and the data proves that. However, based on what we found, the future isn't looking any better, without any significant changes. The ARDI dataset led us to create models that suggested no significant change in the way people have died with alcohol as a main factor of death. This could later suggest that people will continue to die from complications of drinking alcohol, excessively or not. The BRFSS dataset suggested that while the amount of "No's" will still be greater than the "Yes's", the number of "Yes's" still does not change much, meaning we would not be learning from our mistakes, and people will continue to drink alcohol further into the future. Finally, the alcohol consumption dataset showed the change each state has gone through since 1983. We suggest that states like New Hampshire that are rising quickly will still rise, unless some major event occurs (government legislation, etc). Overall, we can see that the impact alcohol has had on the United States will likely not change in the near future.

References

AUC ROC Curve in Machine Learning. GeeksforGeeks. (2024, January 25). https://www.geeksforgeeks.org/auc-roc-curve/

Bernoulli Naive Bayes GeeksforGeeks. (2023, October 25). Bernoulli Naive Bayes - GeeksforGeeks

Centers for Disease Control and Prevention. (n.d.). Alcohol and Public Health: Alcohol-related disease impact (ardi). Centers for Disease Control and Prevention. https://nccd.cdc.gov/DPH_ARDI/default/default.aspx

Centers for Disease Control and Prevention. (n.d.-b). Alcohol-Related-Disease-Impact. Centers for Disease Control and Prevention. https://data.cdc.gov/Alcohol-Related-Disease-Impact/Alcohol-Related-Disease-Impact-ARDI-Application-Ye/jkcv-z5vp/about_data

Centers for Disease Control and Prevention. (n.d.-c). BRFSS: Table of alcohol consumption. Centers for Disease Control and Prevention. https://chronicdata.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-Alcohol-Consumption/dts9-xy2f

F1 Score in Machine Learning. GeeksforGeeks. (2023, December 27). https://www.geeksforgeeks.org/f1-score-in-machine-learning/

Godwin, James Andrew. "Ridge, LASSO, and ElasticNet Regression." *Medium*, Towards Data Science, 16 Apr. 2021, towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3.

ML | Stochastic Gradient Descent (SGD). GeeksforGeeks. (2024, March 14).

https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/

Naive Bayes classifiers. GeeksforGeeks. (2024, March 01). https://www.geeksforgeeks.org/naive-bayes-classifiers/

Roberts, Amber. "Precision: Understanding This Foundational Performance Metric." *Arize AI*, arize, 11 Nov. 2022, arize.com/blog-course/precision-ml/.

U.S. Department of Health and Human Services. (n.d.). Surveillance report . National Institute on Alcohol Abuse and Alcoholism. https://www.niaaa.nih.gov/publications/surveillance-reports/surveillance120