Determining the Usefulness of CDC Data for Disease Risk Prediction:

An exploratory dive into CDC Survey Data

Sara Wirth
Department of Data Science,
University of Colorado
Boulder, CO, USA
sara.wirth@colorado.edu

Srija Vakiti
Department of Data Science,
University of Colorado
Boulder, CO, USA
Srija.Vakiti@colorado.edu

Vaishnavi Asuri
Department of Data Science,
University of Colorado
Boulder, CO, USA
vaas4229@colorado.edu

ABSTRACT

This project aimed to develop a recommender system based on Alcohol Use Disorder (AUD) risk scores for targeted advertisements. However, due to challenges such as data scarcity, the goal was shifted towards evaluating the usefulness of the Centers for Disease Control and Prevention (CDC) data for generating risk scores. The study utilized data scaling and mode imputations, but there is concern that these methods introduced bias, leading to subpar accuracy in the model implementation. The study suggests that a more comprehensive and standardized approach to survey data collection and analysis is necessary for future improvements. The study highlights the critical role of age and dietary habits in the development of heavy drinking habits and emphasizes the need for targeted interventions for populations at higher risk. Future work can include exploring alternative modeling techniques and gathering more comprehensive data to improve the accuracy of risk score predictions and aid in the development of effective interventions for AUD prevention and treatment.

CCS CONCEPTS

Mathematics, computing, theory of computation, information systems, exploratory data analysis

KEYWORDS

Recommender systems, data mining, public health, education, targeted advertising, disease prediction, EDA, CDC, survey, NCHS

ACM Reference format:

Sara Wirth, Srija Vakiti and Vaishnavi Asuri. 2023. Determining the Usefulness of CDC Data for Disease Risk Prediction. CSCI 5502-Data Mining Project. *University of Colorado, Boulder, CO, USA, 2 pages.*

INTRODUCTION

NHIS is a major data collection program of NCHS, CDC that provides current and accurate statistical information on illness and disability in the US. This study aimed to assess the usefulness of NHIS data in predicting individual disease risk, specifically for alcohol use disorder. We utilized data from various questionnaires to model disease risk, which has significant implications in healthcare ranging from product marketing to insurance policies.

AUD is a serious public health issue in the US, causing over 140,000 deaths annually and affecting over 14 million Americans. The CDC survey data can be used to predict AUD risk and target resources and screenings to those at higher risk, potentially mitigating the negative effects of the disorder.

OUR APPROACH

In this section, we first describe the details of the dataset and follow with details about our methods.

The CDC Dataset

NHANES, conducted by the NCHS's DHANES, is an ongoing health and nutrition survey of non-institutionalized civilians in the US. Each year, 5,000 participants of all ages are interviewed at home and examined in a mobile center to collect high-quality data. The survey design has changed to sample larger subgroups of public health interest to improve health status indicator estimates. The 2017-2018 cycle oversampled Hispanic, non-Hispanic black and Asian, low-income non-Hispanic white and other persons, and non-Hispanic white and other persons aged 80 and older.

Questionnaire / Dataset	Number of Variables	Number of People Surveyed
Demographics Data	46	8,704
Alcohol Use Questionnaire	9	5533
Physical Functioning Questionnaire	35	8421
Diet Behavior and Nutrition Questionnaire	45	9254
Drug Use Questionnaire	40	4572
Mental Health-Depression Screener Questionnaire	10	5533

Table (1): The above table describes demographic data across different factors as mentioned in the 'Questionnaire/Dataset' column

All questionnaires and datasets are from NHANES 2017-2018.

RELATED WORK

The major objectives of NHANES are to:

- Estimate the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors;
- Monitor trends in the prevalence, awareness, treatment, and control of selected diseases;
- Monitor trends in risk behaviors and environmental exposures;
- Study the relationship between diet, nutrition, and health;
- Explore emerging public health issues and new technologies; and
- Provide baseline health characteristics that can be linked to mortality data from the National Death Index or other administrative records (e.g., enrollment and claims data from the Centers for Medicare & Medicaid Services).

PROPOSED WORK

NHANES focuses on trends and information gathering. We propose conducting data mining and analysis to determine health risk factors for surveyed individuals. This information could be used by organizations to distribute helpful information to at-risk individuals to improve quality of life. As a proof of concept, we will perform Exploratory Data Analysis (EDA) on survey data from 2017-2018 relating to alcohol and drug use, demographics, physical functioning and activity, nutrition and diet, and depression. We will be develop a pipeline to check the credibility of the dataset to develop risk score. For this we will be implementing six Machine

Learning models and check the results across all of them. They are: Logistic Regression, Support Vector Machine, Decision Trees, Random Forest Classifier, Gradient Boost model and Stacked Classifier. The methodology of conducting them is as mentioned below.

MAIN METHODS

a. Data downloading and conversion

In the below paragraph, it is explained how .xtp files are downloaded and converted to .csv files and combined into DataFrames for pre-processing:

1. Download the data or questionnaire results from the CDC website in their .xtp file format.
2. In the terminal, use the python xport function to convert the .xtp file to .csv file using the following command
    a. python -m xport filename.xtp > filename.csv
3. Upload the .csv
4. The "Variable Descriptions" can be retrieved from the CDC website and manually (or with Python) merged into the .csv data for easier use.
5. In your software, merge the .csv files into a DataFrame for pre-processing
6. Data for answer codes can be found on the CDC website for each questionnaire.

b. Data selection

The large dataset was partitioned into two subsets, and each subset was analyzed using distinct methodologies as part of the present study.

i. *First dataset*

In this report, a dataset consisting of 34 questions focusing on alcohol surveys in relation to age, gender, depression levels, and physical functioning was analyzed. The dataset was filtered to include only the age range of interest and individuals who provided information on their alcohol consumption habits.

Rows with missing data and null values were removed, resulting in a dataset of 4393 individuals. New columns were created to explore the relationship between alcohol consumption, depression, and physical functionality. The data was rescaled to a range of 1-5, and NaN values were replaced with 0 for uniformity.

ii. *Second dataset*

The current study focused on exploring the relationship between alcohol and drug consumption with diet, nutrition, and social and mental health issues. A dataset containing age and gender demographics of participants was used for this purpose.

The initial dataset had 9,200 participants and 85 questions, but 40 unanswered questions were removed. 67 relevant questions were selected, and 15 questions were chosen for analysis from the final dataset of 5,739 participants. Missing values were replaced by mode imputation, and responses were rescaled on a scale of 0 to 5. 'Yes/No' questions were kept as 1 and 2, respectively. This ensured a clean dataset and enabled meaningful conclusions about the relationship between alcohol consumption and physical functionality.

EXPLORATORY DATA ANALYSIS

a. Pre-Risk-Score development EDA

The pie chart generated from the dataset revealed that a significant proportion of the participants either refused to answer or were not aware of the answers to questions related to alcohol and drugs.

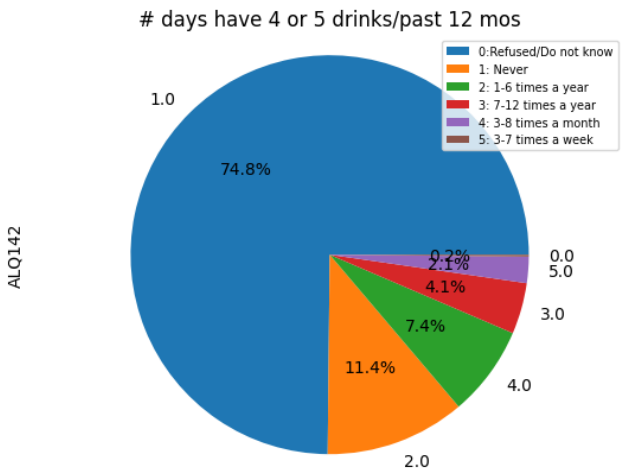

Fig. (1): The percentages of people in each range who answered the question 'Number of days you had 4 or 5 drinks in the past year'

This highlights the potential limitations of self-reported data and the importance of addressing missing or incomplete data in research studies.

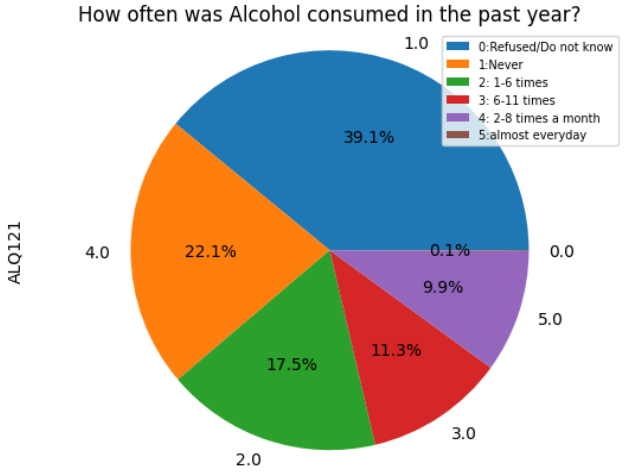

Fig. (2): The percentages of people in each range who answered the question 'How often did you consume alcohol in the past year'

Furthermore, our analysis showed that across all age groups, there was an almost equal distribution of male and female subjects. This balance of gender distribution is a significant advantage in avoiding gender bias and ensuring that the results of the study are representative of the general population.

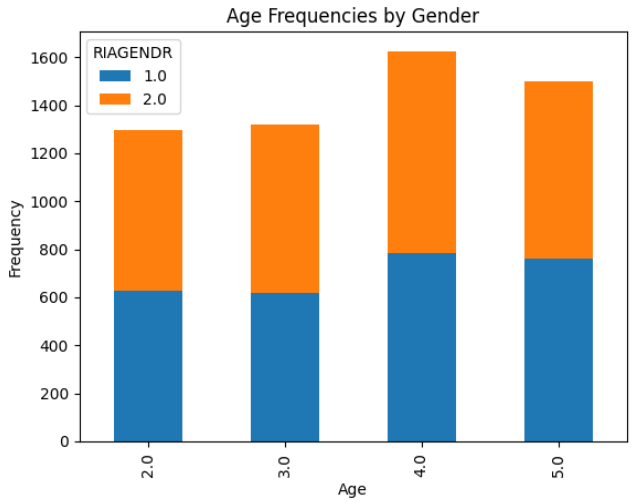

Fig. (3): The ratio of Females(orange) to Males(blue) distributed across 5 age groups ranging from 0 to 90 years

Upon analyzing the first dataset, it was discovered that only 145 individuals out of 4400 participants claimed to be experiencing depression. This low prevalence of depression made the process of data visualization and statistical analysis more complex.

Our analysis indicated a strong correlation among questions related to physical functionality. Specifically, we observed a high degree of correlation between individuals who experienced difficulties with getting out of bed and dressing, and those who had difficulty preparing meals.

This finding suggests that certain physical limitations or challenges may be indicative of other related difficulties in daily activities. The high correlation among these questions related to physical function can also help to identify areas where individuals may require additional support or assistance in daily living.

a. Post-risk-score development EDA

Overall, it was found that the dataset was dominated by those who are at higher risk.

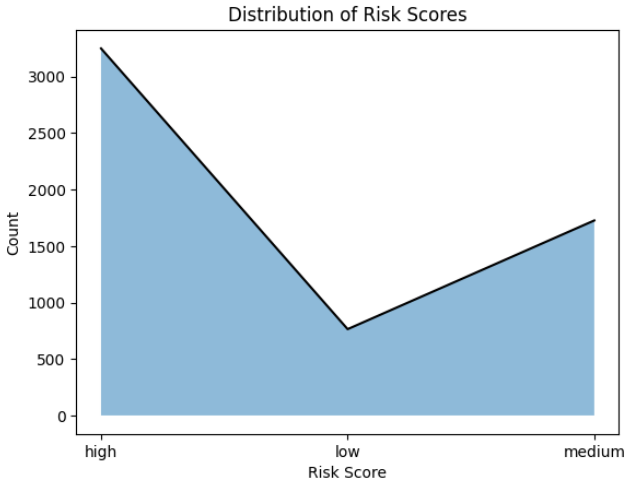

Fig. (4): The distribution of population of three levels of risk scores

Our analysis revealed a surprising finding regarding the probability of developing a habit of a heavy drinking problem. Specifically, we observed that individuals between the ages of 40 and 70 have the highest probability of developing such a habit. However, there is a sharp drop in this probability when it comes to individuals between the ages of 50 and 60.

This finding suggests that age plays a critical role in the development of heavy drinking habits, with individuals in their 40s and 70s being at a higher risk. However, the sharp drop in probability between ages 50 and 60 suggests that there may be other factors at play during this period that could contribute to a decrease in heavy drinking habits. Further research may be needed to explore these factors and better understand the underlying reasons for this trend.

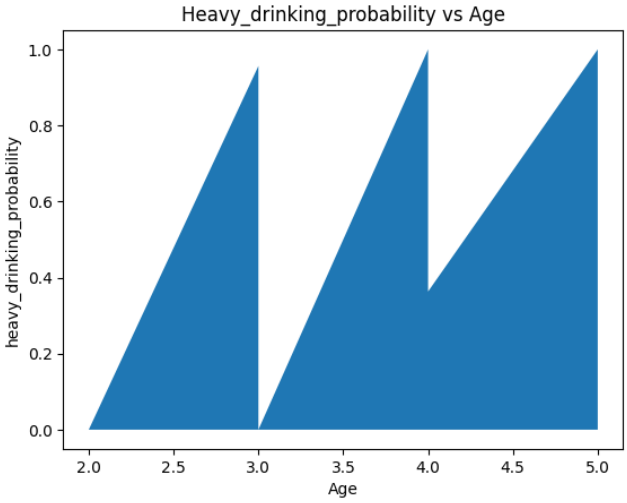

Fig. (5): The probability of developing a heavy drinking habit across different age groups ranging from 0 to 90 years

We found that among those categorized as high and medium risk for developing AUD, women were present in higher numbers across all age groups.

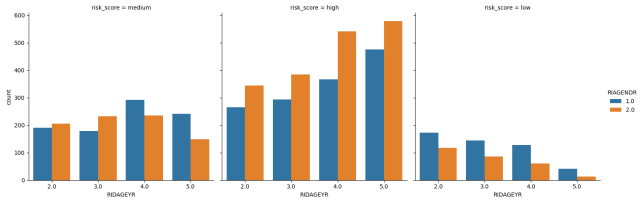

Fig. (6): The ratio of Females(orange) to Males(blue) distributed across 5 age groups in three risk score categories

Furthermore, we observed that individuals who consume homemade food that is mostly ready-to-eat are more prone to developing AUD than those who tend to eat outside. This finding suggests that dietary habits may play a role in the development of AUD and that individuals who tend to consume more processed or ready-to-eat foods may be at a higher risk.

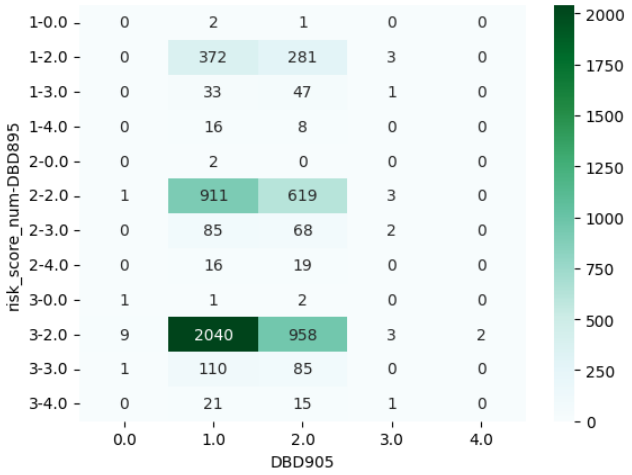

Fig. (7): Contingency table implying a relationship between risk scores and whether or not people eat home-cooked meals

The Ratio of people who answered 'Yes' to experiencing confusion/memory problem and other limitations at work when plotted against Risk scores was found to be higher than those who answered 'No'.

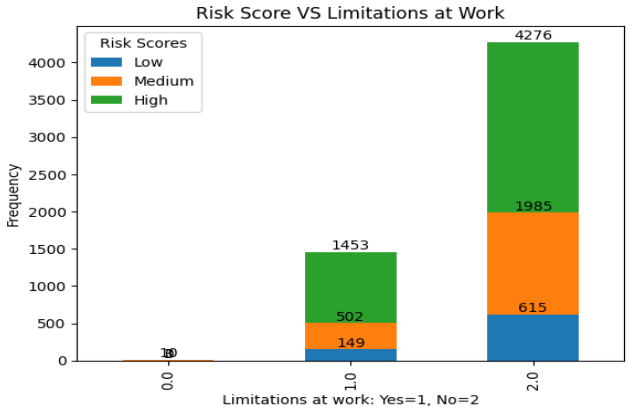

Fig. (8): The ratio of High, Medium and Low risk score populations based on whether or not they face limitations at work

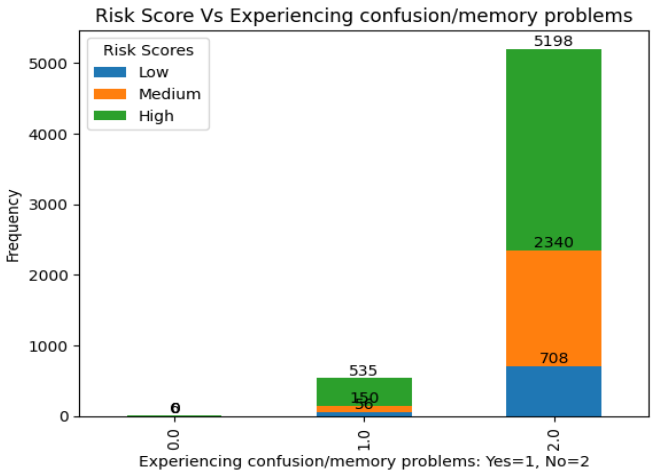

Fig. (9): The ratio of High, Medium and Low risk score populations based on whether or not they experience confusion/memory problems

EVALUATION

This project will be evaluated on the feasibility of mining data from the CDC, our ability to gather usable data, being able to generate risk scores for different health categories, and developing a recommender system that works with this pipeline.

a. On basic KNN based on all features

To gain insights into the performance of each model in the first sub-approach of KNN (which used all the features), a grouped bar chart was plotted to compare their f1, recall, accuracy, and precision values.

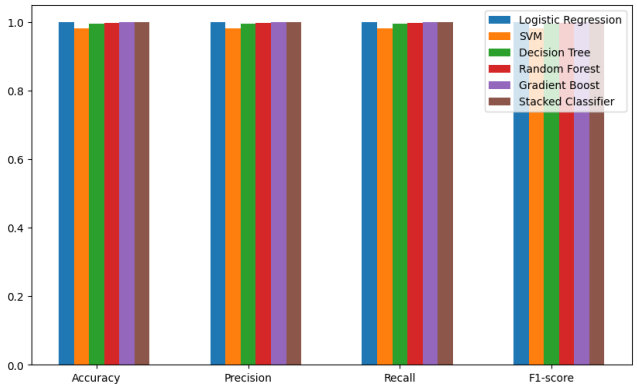

Fig. (10): Evaluation of the four accuracy metrics of all six ML models implemented on basic KNN approach

However, the analysis did not reveal significant differences between the models. Moreover, since all models exhibited high accuracy and perfect confusion matrices, it suggests that they may have overfit the training data, and the results must be interpreted with caution.

b. On t-SNE-based KNN based on 2 features

To gain performance insights of the second sub-approach of KNN (which used t-SNE features), another grouped bar chart was plotted to compare their f1, recall, accuracy, and precision values.

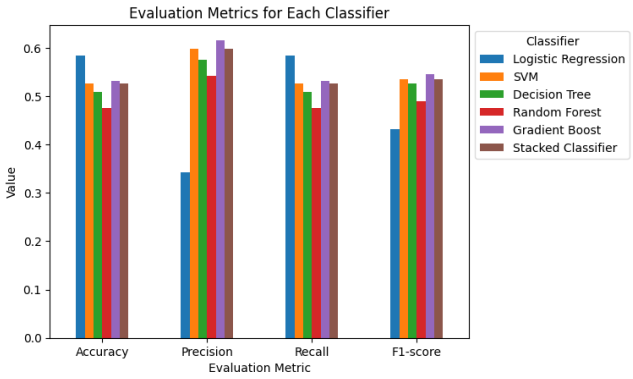

Fig. (11):Evaluation of the four accuracy metrics of all six ML models implemented on t-SNE based approach

Risk Score

a. Risk Score Development

Risk scores can help individuals by providing them with information about their risk level for a particular condition, such as alcohol use disorder (AUD). Developing a risk score helps build targeted advertisements by identifying individuals who are at higher risk for AUD). Once these individuals are identified, targeted ads can be created and delivered to them specifically, with messaging and content that is designed to resonate with their unique characteristics and needs.

This approach provides a quantitative measure for assessing the risk of alcohol use disorder and can be used by public health practitioners and researchers to develop effective prevention and intervention strategies.

b. Approaches

Despite attempting to use both correlation matrix and feature engineering methods to identify important features, neither approach was successful in differentiating between important and unimportant features. As a result, we had to consider all the available features in our analysis.

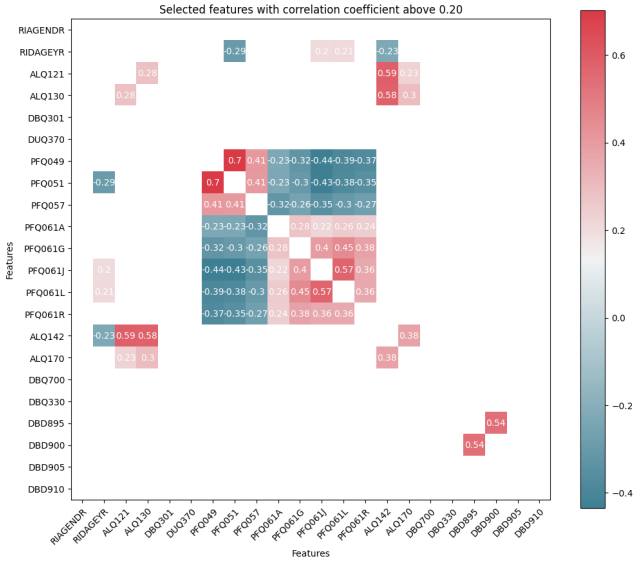

Fig. (12): Confusion matrix that shows some of the lowest and highest correlated features

We decided to generate risk scores based on 2 major approaches–Logistic regression and KNN.

b i) Risk score based on logistic regression

The first approach implemented logistic regression to predict the probability of heavy drinking based on selected numerical features from the processed dataset. The logistic regression model was first trained using the selected features as predictors and the AUD risk score was defined as the target variable.

The resulting model was used to predict the probability of heavy drinking for each individual

in the dataset using the predict_proba() method. The probabilities generated from the model were initially scaled between 0 and 1, and were subsequently rescaled to reduce the number of decimal points. This rescaling process made the output values more manageable and easier to interpret. These probabilities were added as a new column to the original dataset.

b ii) Risk score based on KNN

The second method employed in the study involved clustering the data based on selected features using the KMeans clustering algorithm. This approach was divided into two sub-parts: KMeans clustering using all features and a t-SNE (t-Distributed Stochastic Neighbor Embedding) based KMeans clustering that used two hybrid features.

Basic KNN approach: For the first sub-approach, the features were selected from the processed dataset and assigned to X. The KMeans clustering model was then initialized with 3 clusters, and the model was fit to the data using the fit() function. The predict() function was used to assign cluster labels to each data point. Cluster labels were then assigned risk scores of low, medium, or high using a dictionary, and cluster risk scores were assigned to each data point.

The result was a score for each survey participant ranging between 0 and 2 with 0 as low risk of AUD, and 1 and 2 being medium risk and high risk of AUD respectively. This risk score was appended as a new column to the original dataset.

t-SNE KNN approach: The second sub-approach first imported the t-SNE module from the Scikit-learn library and used it to apply t-SNE to reduce the dimensionality of the data. The number of components was set to 2, and the perplexity was set to 30, while the random state was set to 42.

The importance of each original feature in t-SNE components was also taken into consideration. The perplexity value was first extracted from the t-SNE model parameters. Next, two subplots were created, one for each t-SNE component. The importance of each feature was represented on the y-axis by the KL divergence divided by the logarithm of the perplexity.

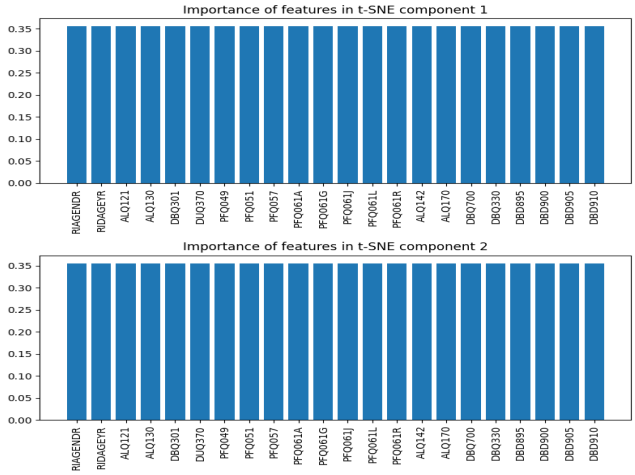

Fig. (13): Feature importance based on t-SNE components

The features in t-sne 1 and 2 had an importance score of 0.35, indicating that they all contributed important information to the data structure. Risk scores were generated for each participant, ranging from 0 to 2, indicating the level of AUD risk. A scatter plot was created using t-SNE components as x and y axes, and cluster labels as point colors. The plot was labeled and included a legend.

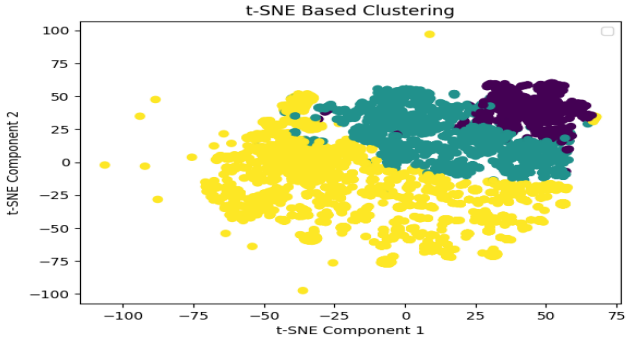

Fig. (14): Clustering based on t-SNE model

MODEL IMPLEMENTATION

a. Implementation on logistic regression results

To better visualize the relationship between the predicted probabilities and the continuous target variable, a scatter plot was created with the predicted probabilities on the x-axis and the actual average drinking quantity on the y-axis.

Each point on the scatter plot represents a survey participant. By comparing the scatter plot with the regression line, it was noted that the points did not scatter randomly around the regression line, but instead showed a clear pattern of deviation from it.

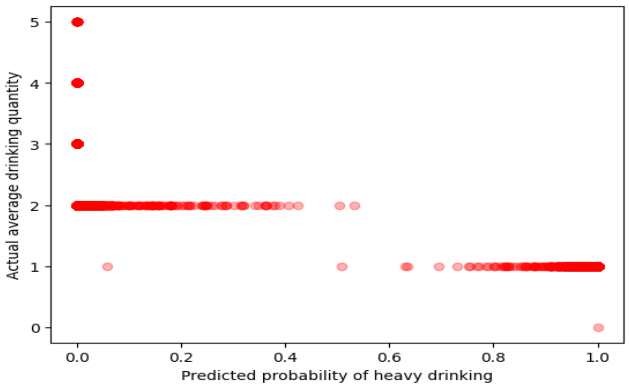

Fig. (15): Predicted probability of heavy drinking Vs actual average drinking

This suggests that the logistic regression model was not accurately predicting the actual values of the target variable. The pattern of deviation indicates that the model may be underestimating or overestimating the actual average drinking quantity for certain groups of participants. This was why KNN based risk score results were considered more apt.

b. Implementation on KNN results

The dataset containing risk scores was split into training and test sets using train_test_split from sklearn.model_selection. Six classifiers were defined using different algorithms, and each classifier was trained on the training set. The accuracy and confusion matrix were calculated on the test set using accuracy_score and confusion_matrix functions from sklearn.metrics. The confusion matrices were visualized using matplotlib.pyplot, with the title indicating the accuracy score, and the x and y axes labeled with predicted and actual values. SVM had an accuracy of 0.993, while the rest had perfect accuracy.

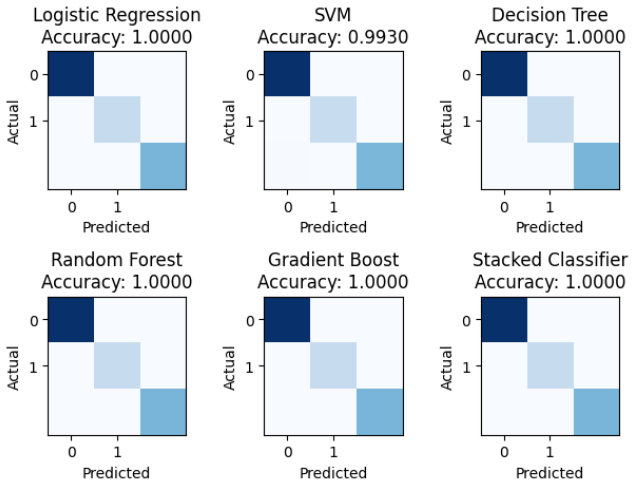

Fig. (16): Confusion matrix based on basic KNN approach

In theory, an accuracy of 1.0 (100%) is possible, indicating that the model has made no incorrect predictions on the test set. However, in practice, achieving such high accuracy is often unlikely. Based on the high accuracy and perfect confusion matrices, it is likely that the model had overfit the training data.

c. Implementation on t-SNE-based KNN results

A new dataset was split into training and test sets and then scaled using the StandardScaler from sklearn.preprocessing. Dimensionality reduction was performed using the t-SNE algorithm from sklearn.manifold to reduce the data to two hybrid dimensions. Six classifiers were then defined and trained using different algorithms, including Logistic Regression, SVM, Decision Tree

Classifier, RandomForest Classifier, Gradient Boosting Classifier, and a Stacking Classifier. The accuracy and confusion matrix were calculated on the test set using the accuracy_score and confusion_matrix functions from sklearn.metrics. The confusion matrices were visualized using matplotlib.pyplot, with darker colors representing higher values.

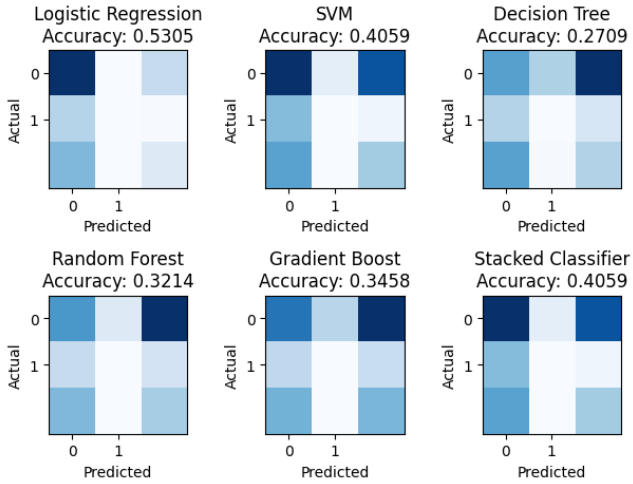

Fig. (17): Confusion matrix based on t-SNE approach

Although the accuracy of the classifiers trained on the t-SNE reduced data was lower than those trained on the original data, t-SNE still provides an improvement over the original results. t-SNE preserves the local structure of the data, resulting in a more meaningful representation of the data in lower dimensions. Although this reduction in accuracy was expected due to information loss during dimensionality reduction, the lower-dimensional representation obtained by t-SNE is more useful for visualization and interpretation, and can also be used as input for other machine learning algorithms.

RESULTS

a. Pre-Risk-Score development

1. In summary, our study revealed that a significant proportion of participants either refused to answer or were not aware of answers related to alcohol and drugs, highlighting the importance of addressing incomplete data in research.
2. Additionally, our analysis showed an equal distribution of male and female subjects across all age groups, minimizing the risk of gender bias.
3. We also found a low prevalence of depression, making data analysis more complex, and identified a strong correlation among questions related to physical functionality, indicating the potential for identifying areas where additional support may be needed.

b. Post-Risk-Score development

1. The dataset had a high proportion of individuals at higher risk, requiring targeted interventions for these populations.
2. Individuals aged 40-70 have the highest probability of developing a habit of heavy drinking, with a sharp drop between ages 50 and 60.
3. Women are present in higher numbers among those categorized as high and medium risk for developing AUD across all age groups.
4. Homemade food that is mostly ready-to-eat increases the risk of developing AUD compared to eating outside, highlighting the need to consider dietary habits in interventions.

c. Model implementation results

1. The logistic regression classifier had the highest accuracy that upon closer examination, was apparent due to its ability to correctly classify instances with low risk scores, as opposed to those with higher risk scores. This was reflected in

its relatively poor precision values, which suggests a high number of false positives.
2. In contrast, Gradient Boost performed well across multiple evaluation metrics, including precision, recall, F1 score, and accuracy. This suggests that it may be a more well-rounded classifier for this particular dataset, with a greater ability to accurately classify instances across a range of risk scores.
3. The SVM classifier had relatively high precision and recall values but slightly lower accuracy and recall. The decision tree and random forest classifiers had relatively low precision values, while the stacked classifier had similar performance to the SVM classifier.
4. Overall, the results suggest that the gradient boosting classifier may be the best choice for this particular dataset, as it had the highest overall performance across all four evaluation metrics.

CHALLENGES

Variable scales: Each variable in the dataset had different scales based on the ranges and values of the survey answers, which made it difficult to compare and analyze them. To overcome this, manual re-scaling of all variables was necessary to ensure that the results obtained were comprehensive and meaningful.

1. Generalizations: Re-scaling variables to much smaller ranges resulted in generalizations, which caused smaller and more nuanced results to go unnoticed. This is a common challenge when working with survey data since there is always a trade-off between granularity and interpretability of the results.
2. Missing values: The data was very inconsistent when it came to missing values, with a significant number of values marked as 'Do not know' or 'Refused to answer.' Additionally, many value points were marked as 'nan' because the data was simply unavailable. Imputation and data manipulation were required to make the data usable, which can introduce bias into the results.
3. Mode imputation: Mode imputation, which was performed to fill in missing values, may not have been the best method for this dataset. This technique caused under-represented populations to remain under-represented, and dominant groups to become even more dominant. This can lead to biased results, especially when analyzing small subgroups.
4. Feature engineering: Some models required us to choose certain features over others, but many of these features had low correlation with each other. Feature engineering was implemented to address this limitation, which involved transforming existing features or creating new ones to improve the model's performance. However, this resulted in sub-par accuracies, which is a common trade-off in machine learning.

CONCLUSIONS

In the beginning our goal was to develop a recommender system based on AUD risk score for targeting advertisements. Due to the challenges faced in the process, mainly lack of data, we had to change our aim to checking the usefulness of CDC data to generate a Risk score.
The original objective was to design a recommender system based on AUD risk score to

target advertisements effectively. However, we were impeded by a lack of data, and consequently, we had to adjust our focus to the evaluation of the utility of CDC data for generating a Risk score.

Although we were successful in creating a standardized scaling system and a robust pipeline, our findings were met with a caveat. Our mode imputations and scaling techniques may have introduced a substantial degree of bias, rendering the usefulness of the results somewhat ambiguous. The model's implementation was also met with subpar accuracy, further emphasizing the limitations of our approach.

However, a negative outcome is still informative and has merit in its own right. If we were to restart our analysis, we would prioritize questions with ample data rather than working with a predetermined goal. This approach would yield a wider range of results, circumventing the need for data manipulation. It is imperative to standardize surveys to improve their efficacy for future analyses.

It is also important to mention here that this project has been a learning experience more than anything. Reproduction of this project without improvisations may not yield desirable or accurate results.

FUTURE WORK

We intend to improve data collection methods to increase sample size and improve data quality
We also realize the importance of exploring different feature engineering techniques, such as PCA or feature selection, to identify the most important variables and improve model accuracy.

We must consider using more advanced machine learning techniques, such as neural networks or ensemble methods, to improve performance. Investigating other risk factors, such as genetics, mental health, and social/environmental influences, to develop more comprehensive models for predicting AUD risk is also important. We also plan to evaluate model generalizability to ensure applicability to broader populations.

Furthermore , we can explore alternative applications of the results, such as public health interventions or personalized treatment plans.

REFERENCES

[1] Parola R, Ganta A, Egol KA, Konda SR. Trauma risk score matching for observational studies in orthopedic trauma dataset and code. Data Brief. 2022 Jan 5;40:107794. doi: 10.1016/j.dib.2022.107794. PMID: 35036491; PMCID: PMC8749164.

[2] SAMHSA, Center for Behavioral Health Statistics and Quality. 2019 National Survey on Drug Use and Health. Table 2.17B – Alcohol Use in Lifetime among Persons Aged 12 or Older, by Age Group and Demographic Characteristics: Percentages, 2018 and 2019. https://www.samhsa.gov/data/sites/default/files/reports/rpt29394/NSDUHD…. Accessed December 8, 2020.

[3] SAMHSA, Center for Behavioral Health Statistics and Quality. 2019 National Survey on Drug Use and Health. Table 2.18B – Alcohol Use in Past Year among Persons Aged 12 or Older, by Age Group and Demographic Characteristics: Percentages, 2018 and 2019. https://www.samhsa.gov/data/sites/default/files/reports/rpt29394/NSDUHD…. Accessed December 8, 2020.

[4] SAMHSA, Center for Behavioral Health Statistics and Quality. 2019 National Survey on Drug Use and Health. Table 2.19B – Alcohol Use in Past Month among Persons Aged 12 or Older, by Age Group and Demographic Characteristics: Percentages, 2018 and 2019. https://www.samhsa.gov/data/sites/default/files/reports/rpt29394/NSDUHD…. Accessed December 8, 2020.

Dataset

Parola, Rown (2021), “STTGMA Matching Dataset”, Mendeley Data, V1, doi: 10.17632/d5b5rx6vmf.1

ANNEXURE

1. SEQN: Sequence number of Respondent
2. RIAGENDR - Gender of the respondent https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm#RIAGENDR

3. RIDAGEYR - Age in years https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm#RIDAGEYR

Alcohol survey questionnaire: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/ALQ_J.htm#ALQ111

4. ALQ111 - Ever had any alcoholic barrage of any kind?
5. ALQ121 - How often did you drink in the past 12 months
6. ALQ130 - Average number of drinks per day in the past 12 months.
7. ALQ142 - Number of days on which the respondent had 4 to 5 drinks each in the past 12 months
8. ALQ270 - Number of times the respondent had 4 to 5 drinks in 2 hours in the past 12 months
9. ALQ280 - Number of times they had more than 8 drinks per day in the past 12 months
10. ALQ290 - Number of times they had more than 12 drinks per day in the past 12 months
11. ALQ151 - Did they ever have more than 4 to 5 drinks per day?
12. ALQ170 - Number of days they had 4 to 5 drinks on the same day on an occasion

Depression survey Questionnaire: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DPQ_J.htm#DPQ010

13. DPQ010 - They have very little interest in doing things
14. DPQ020 - Feeling down, depressed or hopeless
15. DPQ030 - Trouble sleeping or sleeping too much
16. DPQ040 - Feeling tired or having little energy
17. DPQ050 - Poor appetite or over-eating
18. DPQ060 - Feeling bad about themselves
19. DPQ070 - Trouble concentrating on other things
20. DPQ080 - Moving or speaking too slow or too fast
21. DPQ090 - Thought they would be better off dead
22. DPQ100 - Difficulty that these problems have caused

Physical Functionality survey questionnaire: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/PFQ_J.htm#PFQ049

23. PFQ049 - Limitations keeping them from working
24. PFQ051 - Limited in the amount of work they can do
25. PFQ057 - Experiencing confusion or memory problems
26. PFQ059 - Physical, mental, or emotional limitations
27. PFQ061B - Difficulty walking for a quarter mile
28. PFQ061C - Difficulty walking up ten stairs
29. PFQ061D - Difficulty stooping, crouching or kneeling
30. PFQ061E - Difficulty lifting or carrying
31. PFQ061M - Difficulty standing for long periods
32. PFQ061N - Difficulty sitting for long periods
33. PFQ061O - Difficulty reaching up
34. PFQ061A - Difficulty managing money
35. PFQ061J - Difficulty getting in and out of bed
36. PFQ061L- Difficulty dressing themselves
37. PFQ061R - Difficulty attending social events
38. PFQ061G - Difficulty preparing meals
39. DUQ370 - Did they ever use a needle to inject a drug? https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DUQ_J.htm#DUQ370

Diet, Behavior and Nutrition survey questionnaire: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DBQ_J.htm#DBQ010

40. DBQ301 - Number of Community/Givt meals delivered
41. DBQ330 - They eat meals at Community Center
42. DBQ700 - How healthy is their diet?

https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DBQ_J.htm#DBD895

43. DBD895 - Number of meals they eat that are NOT home-prepared
44. DBD910 - Number of frozen meals/pizzas in last 30 days
45. DBD900 - Number of meals from fast food/pizza place
46. DBD905 - Ready to eat foods in the past 30 days