Kaggle Competition Fall 2023

Due Friday December 1st 2023 @ 11:59 PM

Prediction of Alcoholic Status

About Dataset

This dataset is collected from the National Health Insurance Service in Korea. All personal information and sensitive data were excluded.

The purpose of this dataset is to:

1. Analysis of body signal
2. Classification of drinker (Alcoholic status)

This data consists of training and testing data sets; both have 26 predictors.

Training Data: The training data set has 70,000 observations,

Testing Data: (No Response Variable) The testing data set has 30,000 observations.

Introduction:

In this project, we investigate the alcohol drinking dataset. We will first take an initial look at the dataset and use domain knowledge to engineer features. Next, we will perform exploratory data analysis on the dataset and its features. Then, due to the large size of the dataset, we will try to build and predict models to classify the individuals' alcoholic status. Finally, we will evaluate our models on the test set by using classification reports and confusion matrices and accuracies through Kaggle.

More Detailed Feature Table

I will provide the details of the column explanations here, collated with the data descriptions from the dataset description and the original data source, alongside supplementary information referenced from external sources for the more technical variates.

Note 1: 1 dL, or 1 deci-litre, is equivalent to 100 mL (millilitres) or 0.1 L (litres).

Note 2: IUs, or international units, are units of measurement used to quantify the effect/biological activity of a substance. source

Data Features:

```
> names(SA.train)
 [1] "sex"                "age"                "height"
 [4] "weight"             "waistline"          "sight_left"
 [7] "sight_right"        "hear_left"          "hear_right"
[10] "SBP"                "DBP"                "BLDS"
[13] "tot_chole"          "HDL_chole"          "LDL_chole"
[16] "triglyceride"       "hemoglobin"         "urine_protein"
[19] "serum_creatinine"   "SGOT_AST"           "SGOT_ALT"
[22] "gamma_GTP"          "Alchoholic.Status"  "BMI"
[25] "BMI.Category"       "AGE.Category"       "Smoking.Status"
```

1. **sex** represents the sex of the individual - male or female.
2. **Age**: In the original study, age was categorized into 5-year intervals; i.e. 20-24 years, 25-29 years, ..., 85+ years. These intervals were then converted into numeric values by taking the lower bound of the interval. (e.g. 20-24 years -> 20 years)
3. **height** represents the height of the individual, in 5cm increments.
4. **weight** represents the weight of the individual, in 5kg increments.
5. **waist** represents the circumference of the individual's waist.
6. **sight_left** represents the visual acuity of the individual's left eye. Visual acuity measures the ability of the eye to distinguish shapes and object details at a given distance. This ranges from 0.1-2.5, with values <0.1 being shifted up to 0.1. source%20is%20a,detect%20any%20changes%20in%20vision.)
7. **sight_right** represents the visual acuity of the individual's right eye.
8. **hear_left** measures the hearing in the left ear of the individual, with 1 representing it being normal and 2 representing it being abnormal.
9. **hear_right** measures the hearing in the right ear of the individual with the same classification system as hear_left.
10. **SBP** measures the highest systolic blood pressure measured from the individual, in mmHg. Systolic blood pressure measures the pressure in the arteries when the heart beats. source
11. **DBP** measures the diastolic blood pressure measured from the individual, in mmHg. Diastolic blood pressure measures the pressure in the arteries when the heart rests between beats. (Same source as above.)
12. **BLDS** measures the individual's fasting blood glucose, in mg/dL. This represents the concentration of glucose per 100ml of blood prior to eating a meal.
13. **tot_chole** measures the total concentration of (ester and non-ester) cholesterol in the individual, in mg/dL.
14. **HDL_chole** measures the total concentration of cholesterol in the individual's HDL (high-density lipoprotein) region, in mg/dL. HDL cholesterol, also called good cholesterol, absorbs cholesterol in the blood and carries it back to the liver, which then flushes it from the body. Note that higher values of HDL cholesterol can lower the risk of heart disease. source,for%20heart%20disease%20and%20stroke.)
15. **LDL_chole** measures the total concentration of cholesterol in the LDL (low-density lipoprotein) region, in mg/dL. LDL cholesterol, also called bad cholesterol, makes up

most of the body's cholesterol. High levels of this can raise the risk of heart disease and stroke. source

16. **triglyceride** measures the total concentration of triglyceride in the individual's blood, in mg/dL. Triglycerides are a type of lipids (i.e. fat) that circulate in our blood and usually originate from foods we consume. source
17. **hemoglobin** measures the total concentration of hemoglobin in the individual's blood, in g/dL. Hemoglobin is a protein in our red blood cells that carries oxygen. source
18. **urine_protein** measures the amount of protein in the individual's urine. High levels of protein in urine, or proteinuria, can be a sign of many health problems, like heart failure and kidney problems. source. It is unclear what the label encodings (1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)) for this variate means, although we could speculate it refers to the number of standard deviations from the mean the value resides at in the feature distribution.
19. **serum_creatine** measures the concentration of creatinine in the individual's serum (which resides in their blood), in mg/dL. Creatinine is a waste product of creatine, which is produced to supply energy mainly to the muscles. Usually, this is removed by your kidneys entirely; thus, if kidney function is abnormal, the concentration of creatinine will increase. source. Normal values are around 0.8-2.7 mg/dL.
20. **SGOT_AST** measures the SGOT (Glutamate-oxaloacetate transaminase) - AST (Aspartate transaminase) value in IU/L, which are values in blood tests that quantify liver, heart and other organs' performance. In particular, when these are damaged, the value of this goes up. Normal values are around 0-40 IU/L.
21. **SGOT_ALT** measures the SGOT (Glutamate-oxaloacetate transaminase) - ALT (Alanine transaminase) value in IU/L, which are values in blood tests that quantify liver performance. In particular, when these are damaged, the value of this goes up. Normal values are around 0-40 IU/L.
22. **gamma_GTP** measures the gamma-GTP (y-glutamyl transpeptidase) value in IU/L, which are values that quantify liver function in the bile dict. Normal values of this are around 11-63 IU/L for men and 8-35 IU/L for women.
23. **Alcoholic.Status**: The response variable is a flag that indicates whether the individual is an alcoholic or not with Y: Yes and N: No levels.
24. **BMI**: Body Mass Index ranges from 12 – 42.5
25. **BMI.Category**: BMI.Category has four levels: Underweight, Healthy, Obese, and Overweight.
26. **AGE.Category**: AGE.Category has four levels: Young, Mid-aged, Old, and Very Old
27. **Smoking.Status** measures the individual's smoking state, with three levels: “never smoked, “used to smoke but quit”, and “still smoking”.

Some Facts Regarding Alcohol Consumption in the U.S.

- In 2021, 60% of U.S. adults drank alcoholic beverages, which decreased from 65% as reported in 2019. The average number of drinks consumed in the previous seven days also decreased.
- According to the 2019 National Survey on Drug Use and Health (NSDUH), 85.6 percent of people ages 18 or older reported that they drank alcohol at some point in their lifetime; 69.5 percent reported that they drank in the past year; 54.9 percent reported that they drank in the past month.

- Globally, the United States ranks 25th for alcohol consumption, with about 8.7 liters of pure alcohol consumed per person per year. This is above the global average of 8.3 liters.
- According to the 2021 NSDUH, 213.2 million adults ages 18 and older (84.0% in this age group) reported that they drank alcohol at some point in their lifetime.
- According to data published by the National Institute on Alcohol Abuse and Alcoholism, U.S. residents' consumption of alcoholic beverages has been on a steady incline since 1995, with a whopping 7.9 billion gallons consumed in 2020.

Alcohol Problems in America include:

- Over half of all American adults have a family history of problem drinking or alcohol addiction.
- More than 10% of U.S. children live with a parent with alcohol problems.
- An estimated 88,000 people die from alcohol-related causes annually.
- Alcohol is the third-leading cause of preventable death in the United States.
- Excessive alcohol use is associated with an increased risk of injuries, chronic diseases such as liver disease and heart disease, several cancers, and poor pregnancy outcomes.
- Excessive alcohol use is responsible for more than 140,000 deaths in the United States each year and it shortens the lives of those who die by an average of 26 years. Excessive alcohol use was associated with an economic cost of $249 billion in 2010.
- Recent estimates by the Centers for Disease Control and Prevention (CDC) attribute more than 95,000 deaths per year to alcohol in the U.S. alone. That's 261 deaths every day. Many of these are related to the effects of long-term alcohol use: liver disease, heart disease and certain cancers. Aug. 8, 2023

Short-Term Health Risks

Excessive alcohol use has immediate effects that increase the risk of many harmful health conditions. These are most often the result of binge drinking and include the following:

- Injuries, such as motor vehicle crashes, falls, drownings, and burns.
- Violence, including homicide, suicide, sexual assault, and intimate partner violence.
- Alcohol poisoning, a medical emergency that results from high blood alcohol levels.
- Risky sexual behaviors, including unprotected sex or sex with multiple partners. These behaviors can result in unintended pregnancy or sexually transmitted diseases, including HIV.
- Miscarriage and stillbirth or fetal alcohol spectrum disorders (FASDs) among pregnant women.
- Ride-sharing has decreased alcohol-related US traffic fatalities by 6.1% and reduced overall US traffic deaths by 4%. (National Bureau of Economic Research)
- As of 2023, every day, 37 people in the United States die in car crashes with an alcohol-impaired driver. This is one death every 39 minutes. (CDC)

- More than a quarter (31%) of all traffic-related deaths are the direct result of alcohol impairment. (NHTSA)
- There were 13,384 deaths from drunk driving crashes in 2021. (NHTSA)
- Over the 10-year period from 2011 to 2021, there were on average 11,000 deaths a year due to drunk driving.
- 1 in 4 crashes with teens involve an underage drunk driver (Mothers Against Drunk Driving)
- Drivers with a Blood Alcohol Content of over 0.10 are 7 times more likely to be involved in a fatal accident than sober drivers. (Responsibility.org)
- Over 10,000 Americans a year are killed by drunk drivers, about 1,000 of them being children. (CDC)

Long-Term Health Risks

Over time, excessive alcohol use can lead to the development of chronic diseases and other serious problems including:

- High blood pressure, heart disease, stroke, liver disease, and digestive problems.
- Cancer of the breast, mouth, throat, esophagus, voice box, liver, colon, and rectum.
- Weakening of the immune system, increasing the chances of getting sick.
- Learning and memory problems, including dementia and poor school performance.
- Mental health problems, including depression and anxiety.
- Social problems, including family problems, job-related problems, and unemployment.
- Alcohol use disorders, or alcohol dependence.

By not drinking too much, you can reduce the risk of these short- and long-term health risks.

Smoking

- Worldwide, tobacco smoking (including second-hand smoke) was the second-leading risk of mortality and contributed to an estimated 8.7 million deaths in 2019. In 2019, smoking ranked third in causing global disability-adjusted life years (DALYs).
- In the US, tobacco use was the second leading risk factor for death and the leading cause of DALYs in 2016.
- A meta-analysis of 23 prospective and 17 case-control studies of cardiovascular risks associated with secondhand smoke exposure demonstrated 18%, 23%, 23%, and 29% increased risks for total mortality, total CVD, CHD, and stroke, respectively, in those exposed to secondhand smoke.
- Tobacco use is one of the leading preventable causes of deaths in the US and globally.
- According to a 2013 study, overall mortality among US smokers was 3 times higher than that for never smokers.
- In 2019, 31.2% of high school students and 12.5% of middle school students used any tobacco products. Additionally, 5.8% of high school students and 2.3% of middle school students smoked cigarettes in the past 30 days.
- In 2018, 13.7% of adults were current smokers (15.6% of males and 12.0% of females)
- Among adults in 2018, 22.6% of American Indians or Alaska Native adults, 14.6% of NH Black adults, 7.1% of NH Asian adults, 9.8% of Hispanic adults, and 15.0% of NH White adults were current smokers.

Physical Inactivity

- In 2018, 25.4% of adults did not engage in leisure-time physical activity.
- In 2018, the overall prevalence of meeting the 2018 Physical Activity Guidelines for Americans for both aerobic and muscle-strengthening guidelines was 24.0% in adults (NH White, 25.7%; NH Black 19.9%; Hispanic or Latino, 21.4%; Asian 22.9%; American Indian/Alaska Native, 19.1%.
- Among students in grades 9-12 in 2017, only about 26.1% met the AHA recommendation of 60 minutes of exercise every day. More high school boys than girls reported having been physically active at least 60 minutes per day on all 7 days.

Nutrition

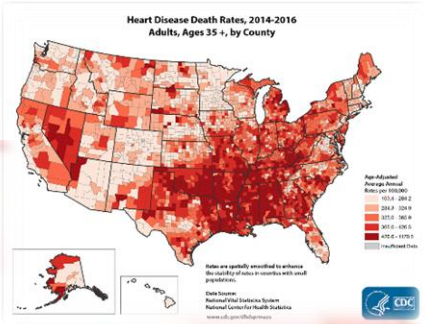
- Between 2003 to 2004 and 2015 to 2016 in the United States, the mean AHA healthy diet score improved in adults. The prevalence of a poor diet improved from 56.0% to 47.8% for the primary score and 43.7% to 36.4% for the secondary score.
    - Changes in score were largely attributable to increased consumption of whole grains and nuts, seeds, and legumes and decreased consumption of SSBs. No significant changes were observed for consumption of total fruits and vegetables, fish and shellfish, sodium, processed meat, and saturated fat.
- Similar changes in AHA healthy diet scores between 2003 to 2004 and 2015 to 2016 were seen in minority groups and those with lower income or education, although significant disparities persisted. The proportion with a poor diet decreased from 64.7% to 58.3% for NH Black individuals, from 66.0% to 57.5% for Mexican American individuals, and from 54.0% to 45.9% for NH White individuals. The proportion with a poor diet (<40% adherence) decreased from 50.7% to 38.8% in adults with income-to-poverty ratio ≥3.0, but only from 67.7% to 59.7% in adults with income-to-poverty ratio <1.3.

Overweight/Obesity

- In the US, the prevalence of obesity among adults increased from 1999 to 2000 through 2017 to 2018 from 30.5% to 42.4%.
- In the US between 2015 and 2018, the prevalence of overweight and obesity among children and adolescents age 2-19 years, was 35.4% (16.4% were overweight and 19.0% were obese).
- According to 2015 to 2016 data, the prevalence of obesity for children aged 2 to 5 years was 13.9 %; for children aged 6 to 11 years, prevalence was 18.4%; and for adolescents aged 12 to 19 years, prevalence was 20.6 %.
- Worldwide, between 1980 and 2013, the proportion of overweight or obese adults increased from 28.8% to 36.9% among males and from 29.8% to 38.0% among females.
- According to the Global Burden of Disease 2019 study, age-standardized mortality rates attributable to high BMI are generally lower in high-income Asia Pacific, Western Europe, East Asia, Australasia, and South Asia.

Cholesterol

- Using data from 2015 to 2018, 93.9 million, or 38.1%, of US adults had total cholesterol of 200 mg/dL or higher. The race and gender breakdown was:
  - 35.0% of NH White males
  - 41.8% of NH White females
  - 31.0% of NH Black males
  - 33.4% of NH Black females
  - 37.7% of Hispanic males
  - 37.3% of Hispanic females
  - 38.6% of NH Asian males
  - 38.6% of NH Asian females



- Using data from 2015-2018 about 28.0 million, or 11.5%, of US adults had total cholesterol of 240 mg/dL or higher. The race and gender breakdown were:
  - 10.1% of NH White males
  - 13.1% of NH White females
  - 9.2% of NH Black males
  - 10.5% of NH Black females
  - 12.4% of Hispanic males
  - 9.2% of Hispanic females
  - 13.0% of NH Asian males
  - 10.3% of NH Asian females

- Using data from 2013 to 2016, 28.9% of American adults had high levels of LDL cholesterol (the "bad" kind; 130 mg/dL or higher).
- Using data from 2015 to 2018, 17.2% of American adults had low levels of HDL cholesterol (the "good" kind; less than 40 mg/dL).

High Blood Pressure (HBP)

- Using data from 2015 to 2018, 47.3% of US adults had hypertension.
- In 2018, there were 95,876 deaths primarily attributable to HBP.
- In 2018, the age-adjusted death rate primarily attributable to HBP was 24.0 per 100,000.

Copyrights:

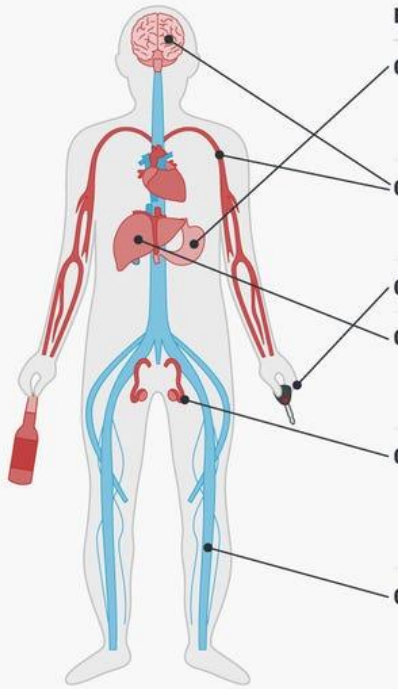
References

1. Yuan, H., An, J., Zhang, Q., Zhang, X., Sun, M., Fan, T., Cheng, Y., Wei, M., Tse, G., Waintraub, X., Li, Y., Day, J. D., Gao, F., Luo, G., & Li, G. (2020). Rates and Anticoagulation Treatment of Known Atrial Fibrillation in Patients with Acute Ischemic Stroke: A Real-World Study. Advances in therapy, 37(10), 4370–4380. https://doi.org/10.1007/s12325-020-01469-w
2. Betts, K. A., Hurley, D., Song, J., Sajeev, G., Guo, J., Du, E. X., Paschoalin, M., & Wu, E. Q. (2017). Real-World Outcomes of Acute Ischemic Stroke Treatment with Intravenous Recombinant Tissue Plasminogen Activator. Journal of stroke and

cerebrovascular diseases: the official journal of National Stroke Association, 26(9), 1996–2003. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.06.010
3. Abzhandadze, Tamar & Reinholdsson, Malin & Sunnerhagen, Katharina. (2020). NIHSS is not enough for cognitive screening in acute stroke: A cross-sectional, retrospective study. Scientific Reports. 10. 10.1038/s41598-019-57316-8.
4. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3, 160035. https://doi.org/10.1038/sdata.2016.35
5. Woodfield, R., Grant, I., UK Biobank Stroke Outcomes Group, UK Biobank Follow-Up and Outcomes Working Group, & Sudlow, C. L. (2015). Accuracy of Electronic Health Record Data for Identifying Stroke Cases in Large-Scale Epidemiological Studies: A Systematic Review from the UK Biobank Stroke Outcomes Group. PloS one, 10(10), e0140533. https://doi.org/10.1371/journal.pone.0140533
6. Mitchell, J & Collen, Jacob & Petteys, S & Holley, Aaron. (2011). A simple reminder system improves venous thromboembolism prophylaxis rates and reduces thrombotic events for hospitalized patients. Journal of thrombosis and haemostasis : JTH. 10. 236-43. 10.1111/j.1538-7836.2011.04599.x.
7. Stenetorp, P & Pyysalo, Sampo & Topic, Goran & Ohta, Tomoko & Ananiadou, Sophia & Tsujii, Jun'ichi. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. The 3th Conference of the European Chapter of the Association for Computational Linguistics; Avignon, France. 102-107.
8. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37-46. doi:10.1177/001316446002000104
9. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.
10. Code for extracting NIHSS scores from MIMIC-III. GitHub. https://github.com/huangxiaoshuo/NIHSS_IE [Accessed: 19 January 2021]
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069781/#:~:text=Visual%20acuity%20(VA
12. https://www.cdc.gov/bloodpressure/about.htm#:~:text=Blood%20pressure%20is%20measured%20using,your%20heart%20rests%20between%20beats.
13. https://www.cdc.gov/cholesterol/ldl_hdl.htm#:~:text=HDL%20(high%2Ddensity%20lipoprotein
14. https://www.nhlbi.nih.gov/health/high-blood-triglycerides#:~:text=Triglycerides%20are%20a%20type%20of,does%20not%20need%20right%20away.
15. https://www.mountsinai.org/health-library/tests/protein-urine-test
16. https://www.mountsinai.org/health-library/tests/creatinine-blood-test#:~:text=Creatinine%20is%20a%20chemical%20waste,body%20entirely%20by%20the%20kidneys.
17. Drinking too much alcohol can harm your health. Learn the facts | CDC

Kaggle Competition Link:
https://www.kaggle.com/t/6ca3fbd818b84ba0b40647b0609ad9c4

HOW ALCOHOL AFFECTS YOUR BRAIN/BODY AS YOU DRINK



BAC	EFFECTS
0.0	Your body can process ~1 drink an hour. Alcohol starts to enter your **bloodstream** both immediately through your **stomach lining** and more slowly after passing through your stomach and intestines.
0.03	You start to feel warmer as **blood vessels** dilate and intensifies mood and emotion by increasing the release of serotonin and endorphins in the brain.
0.08	By this point, you are legally too intoxicated to **drive.**
0.12	Coordination and mental judgment are impaired. You may slur their speech. Consistently stressing your **liver** can cause fat or scar tissue to build up, impairing its function over time.
0.2	You will most likely appear very drunk and have lost significant motor control and mental function. Many people lose the **ability to perform sexually**, mostly because alcohol disrupts neurotransmitters in the brain. This also causes poor sleep.
0.3	You could lose consciousness and your central **nervous system** becomes depressed. Higher BAC levels are associated with an increased risk of alcohol poisoning and death.

NOTE: Factors including individual variation, gender, physiology (i.e. – presence of food, metabolism, fitness, genetics and drug use) might affect how you process alcohol. Some people with a high BAC may appear to be much more sober than they are.

SOURCES: NIAAA.gov; Johns Hopkins University; Brown University Wellness Center

BUSINESS INSIDER