UNIVALI - UNIVERSIDADE DO VALE DO ITAJAÍ
CURSO: Mestrado em Computação Aplicada
DISCIPLINA: Análise Exploratória de Dados
ALUNO: Khalil G. Q. de Santana

Alcoolismo e Vida Acadêmica

Preâmbulo: A apresentação deste relatório está disponível neste link: https://youtu.be/AEFKP1KB_7A . Demais artefatos como código-fonte e *dataset* utilizado estão neste https://github.com/KhalilSantana/AED-TFinal.

Introdução

O consumo de substâncias psico-ativas como o álcool é um hábito social, cultural, sendo seu uso observado em inúmeras sociedades ao redor do mundo. O consumo moderado desta substância é uma fonte de lazer para muitos[3][4][8] , contudo a mesma também se encontra relacionada com uma miríade de problemas, desde acidentes automotivos ao câncer[9].

Dentro desta perspectiva, seria o consumo alcoólico também relacionado ao desempenho acadêmico? Apesar da simplicidade desta preposição (hipótese), a mesma carece de uma análise cuidadosa para sua verificação. Sendo assim, este trabalho visa estudar a relação do consumo alcoólico e o desempenho acadêmico por meio da análise de *datasets* sobre o tema.

Desenvolvimento & Metodologia

Dataset selecionado

O conjunto de dado selecionado foi obtido por meio da plataforma de ciência de dados Kaggle. Uma pesquisa com as palavras-chave "*alchohol*" e "*student*" resultou em 30 resultados, contudo múltiplos destes são 'remixagens' do *dataset* original, em titulado "*Student Alcohol Consumption*". Este dataset foi então superficialmente analisado visando verificar:

- A procedência dos dados (referencial bibliográfico, local e intervalo da coleta).
- A qualidade e documentação do *dataset,* como legendas para cada atributo e ausência de defeitos como deslocamento de colunas ou similares.
- Sua capacidade de comprovar (ou rejeitar) a hipótese previamente estabelecida por meio dos atributos contidos

Esta análise superficial foi favorável, e, portanto, o *dataset* foi selecionado e um fichamento do mesmo pode ser observado no Quadro 1. Em resumo, se trata de um conjunto de dados obtido por meio de um questionário de 37 perguntas aplicado a 788 alunos de duas escolas em Portugal durante 2005 e 2006.

Dataset Selecionado	
Título	*Student Alcohol Consumption: Social, gender and study data from secondary school students*
URL	https://www.kaggle.com/datasets/uciml/student-alcohol-consumpt

	ion
Artigo	P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
Fonte	Escolas do ensino médio. Em específico, a escola Gabriel Pereira e Mousinho da Silveira em Portugal [2]
Intervalo	2005-2006
Nº Indivíduos	788
Anonimizado	Sim
Nº Atributos	33 colunas
Nº Entradas	1044 linhas, distribuídas entre dois conjuntos (disciplina de matemática e português). Sendo 382 alunos presentes em ambas disciplinas.
Formato	CSV, codificado como UTF-8
Tamanho	Aproximadamente 20KB (comprimido), ou 110KB descomprimidos.
Data de Acesso	27/06/2023

Quadro 1: Fichamento do *dataset* selecionado

É importante salientar que este conjunto de dados já se encontra anonimizado, fato que toma especial importância ao considerar a sensibilidade do tópico em questão, assim como a idade dos indivíduos questionados (15-22 anos). Ademais, idade mínima observada pode surpreender, contudo, é importante considerar que fatores culturais e legais no país em questão são distintos e certas bebidas alcoólicas eram permitidas para menores de 18 anos até 2013 [6], enquanto os dados coletados precedem esta mudança.

Materiais e Métodos

O conjunto de dados em questão foi analisado utilizando a linguagem de programação Python, por meio da plataforma de computação em nuvem Google Collab. Ademais, bibliotecas renomadas no campo de ciência de dados foram utilizadas:

- Pandas: Provendo o *DataFrame*, a principal estrutura de manipulação dos dados utilizada.
- Scikit (Scipy): Provendo implementações de funções estatísticas diversas, como normalização e correlação.
- Matplotlib & Seaborn: Possibilitam a visualização dos dados por meio de gráficos como histogramas, gráficos de linha, dentre outros.

Tais ferramentas foram adotados pela sua alta qualidade, a aptidão das mesmas para o problema em questão, assim como a familiaridade do autor com as mesmas.

Tendo em vista que a hipótese deste trabalho não distingue entre as disciplinas, ambos conjuntos (CSVs) foram concatenados após confirmar que seus atributos (colunas) eram idênticos e estavam ordenados da mesma maneira. Da mesma maneira, o atributo que distingue a escola em questão foi removido já que o mesmo também não é relevante para o problema proposto.

Outro a escolha metodológica importante é tratamento do consumo alcoólico, uma vez que esta base de dados distingue entre consumo em finais de semana e durante a semana, este ultimo foi selecionado como o atributo para a verificação da hipótese. O grau de significância foi estabelecido como α=0,05 (5%). Por fim, a análise do sucesso ou fracasso dos alunos se limita a nota final, isto é, foram desconsideradas as notas intermediárias.

Análise Exploratória de Dados

Esta seção visa analisar o *dataset* selecionado, partindo de verificações de consistência e então realizando análises uni variadas e multi-variadas.

Passo 1: Obter os dados

Os dados foram baixados por meio da interface Web da plataforma Kaggle[1], tentativas de automatizar este processo foram frustradas devido a *backend* da mesma não possuir uma URL estática para o *dataset*, e sim URLs que expiram. Após obter o *dataset* o mesmo foi examinado e arquivos de interesse foram enumerados: (i) 'student-mat.csv' e (ii) 'student-por.csv', contendo, respectivamente, os dados pertinentes a disciplina de matemática e português.

Passo 2: Importar os dados

O .ZIP do *dataset* foi lido e descomprimido em uma pasta temporária, seguido da leitura dos ambos CSVs em *DataFrames* do Pandas, os quais foram concatenados utilizando a função pd.concat(), como ilustrado no Quadro 2.

```
dataset_file = '/Mestrado/EAD/AED-TrabalhoFinal/dataset.zip'
work_dir = '/tmp/'
with zipfile.ZipFile(dataset_file, 'r') as zip_ref:
    zip_ref.extractall(work_dir)
df_math = pd.read_csv('/tmp/student-mat.csv')
df_portuguese = pd.read_csv('/tmp/student-por.csv')
df = pd.concat([df_math, df_portuguese])
```

Quadro 2 - Leitura dos dados

Passo 3: Explorar a estrutura dos dados

Em seguida a estrutura do *dataset* foi examinada por meio dos métodos do DataFrame como `shape`, `columns` e `.describe()`, obtendo um sumário do formato dos dados, o cabeçalho (nome) de cada coluna e um sumário dos atributos, respectivamente. Todos os atributos e legenda dos contidos no *dataset* estão dispostos no Apêndice 1. Em resumo, o *dataset* inclui 1044 linhas e 33 colunas, as quais incluem atributos discretos (categorias de empregos, idades, etc) e contínuos (notas).

Passo 4: Limpeza do *dataset*

Durante esta análise não foram identificadas colunas nulas ou qualquer outro motivo para que entradas ou atributos sejam imputados.

Passo 5: Análise Uni variada

Tendo em vista o número de atributos contidos neste conjunto de dados, a análise uni variada nesta seção limita aos atributos de interesse a hipótese, isto é, o consumo alcoólico e notas finais.

Notas finais

De acordo com [5], as notas durante o ensino médio em Portugal são pontuadas zero à vinte, onde uma nota igual ou superior a dez é suficiente para aprovação. A distribuição das notas finais pode ser observada em Figura 1 e a porcentagem de aprovações na Figura 2, ademais, a Tabela 1 sumariza os valores mínimos, máximos, média e desvio padrão deste atributo.

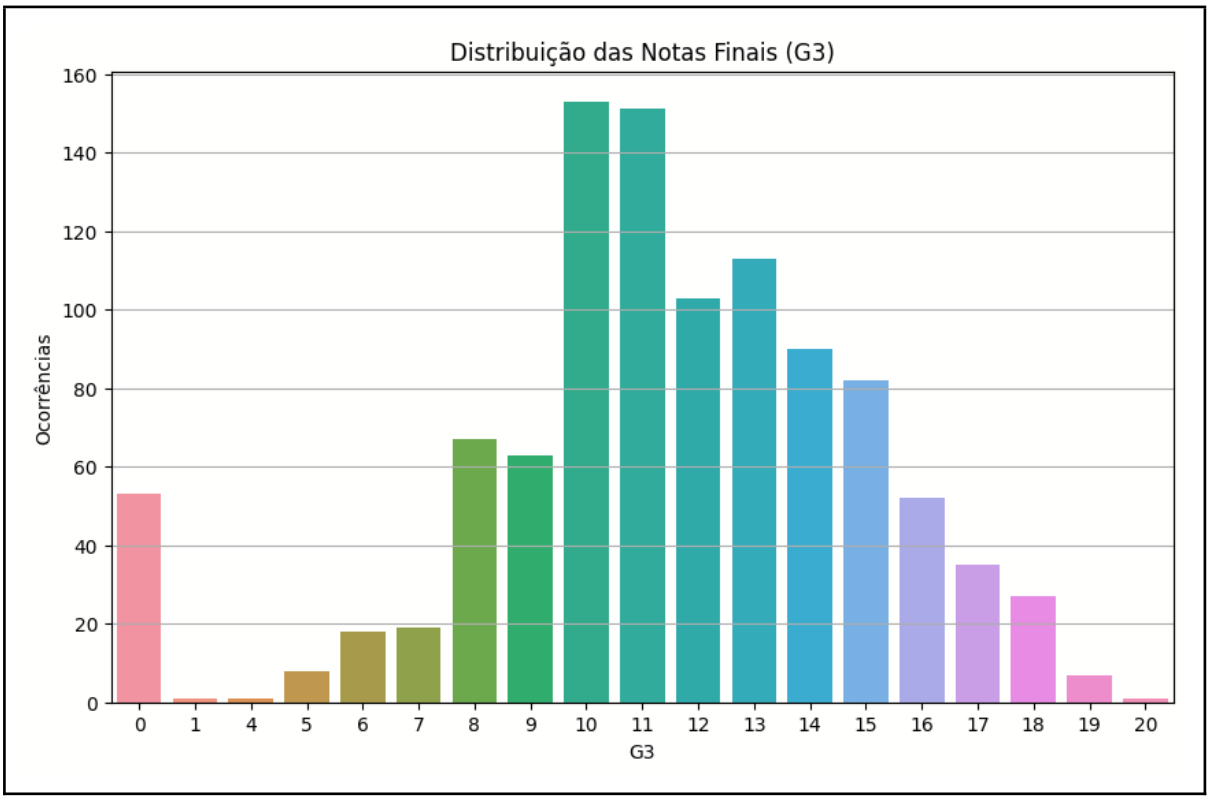
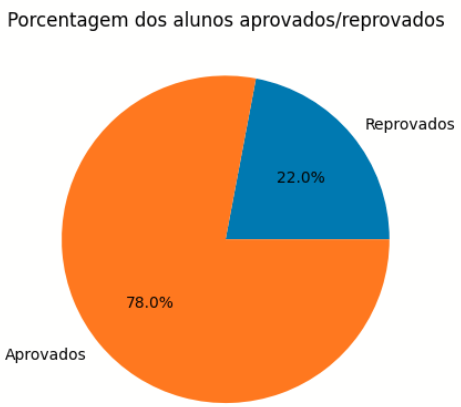


Figura 1: Distribuição de notas finais (G3) de ambas as classes

Nome	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**
G3 (Nota Final)	0.0	20	11.34	3.86

Tabela 1: Métricas estatísticas sobre a nota final



Como é possível observar na Figura 1, existe uma semelhança entre a distribuição de notas e uma distribuição Gaussiana. Contudo, há picos ou desvios notáveis para as notas dez e onze, assim como a nota zero. Foi estipulado que esta (notas zero) estão relacionadas a absências ou desistências, contudo esta hipótese não foi estudada.

Figura 2: Proporção de aprovação/reprovação

Consumo Alcoólico

Apesar do consumo alcoólico durante o dia-a-dia ser o atributo maior de interesse, por completude e como ponto de contraste o consumo alcoólico durante finais de semana também foi incluso nesta seção. A Figura 3 e 4 ilustram, respectivamente, o consumo alcoólico durante finais de semana e durante o dia-a-dia.

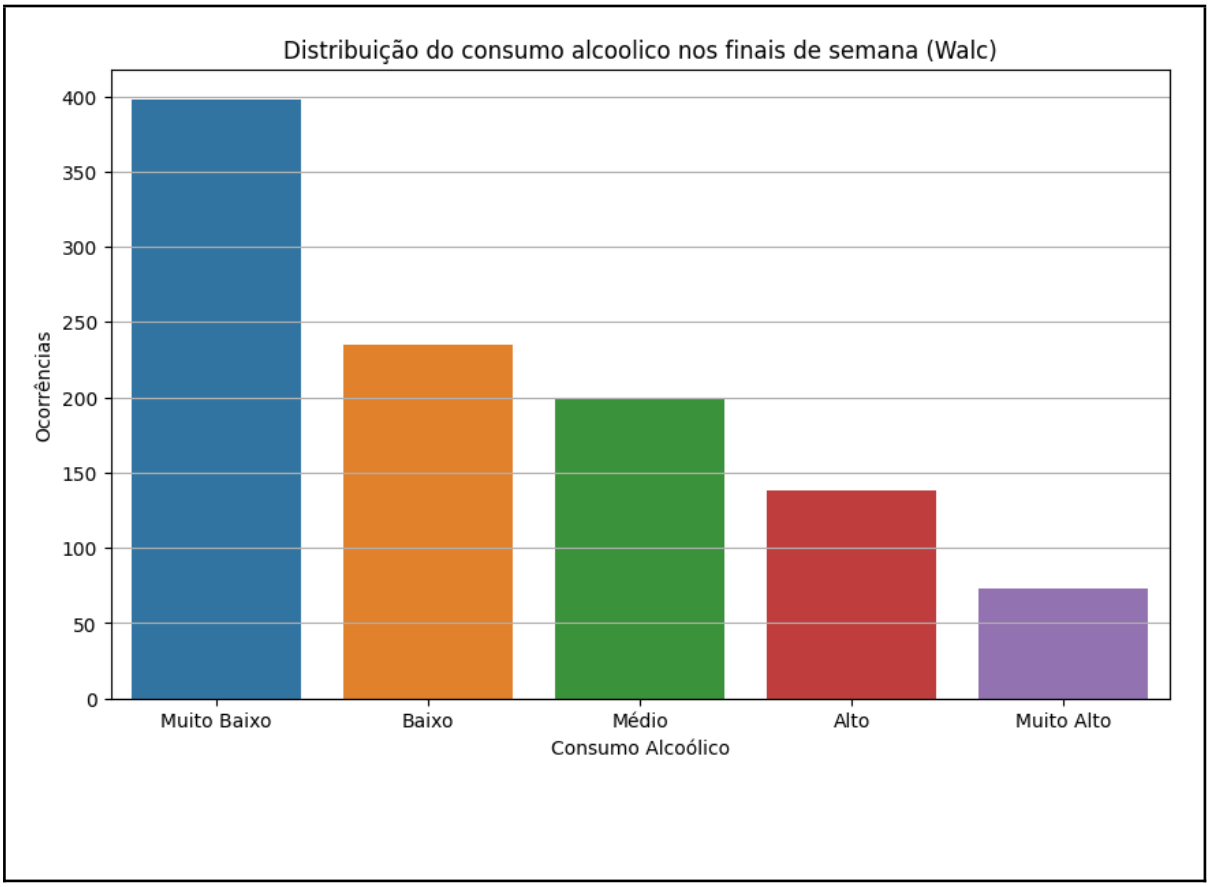


Figura 3: Consumo alcoólico durante finais de semana

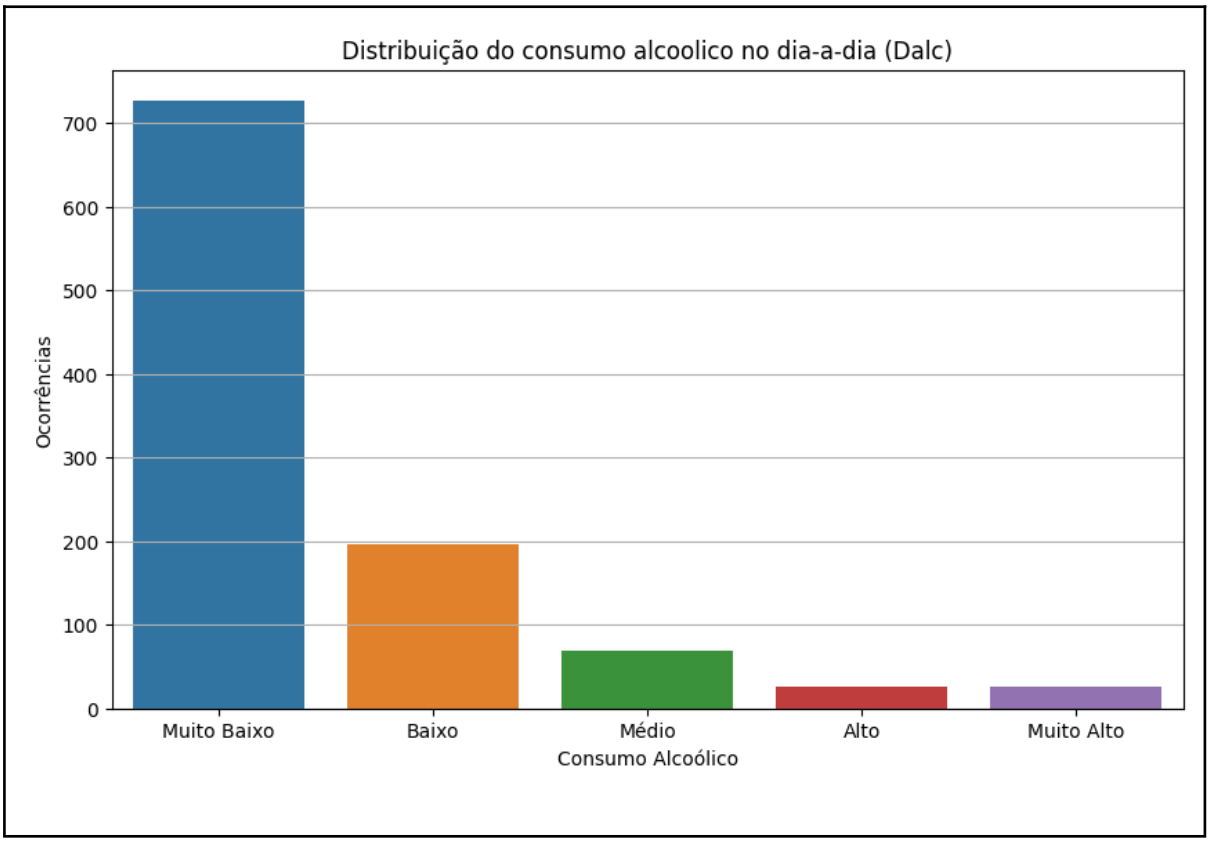


Figura 4: Consumo alcoólico durante o dia-a-dia (Dalc)

Passo 6: Análise Multi variada

Esta seção descreve relações entre múltiplos atributos, em específico, a nota final (G3) e o consumo alcoólico diário (Dalc) e durante finais de semana (Walc). Em figura 5 é possível observar um mapa de calor entre os atributos numéricos deste dataset.

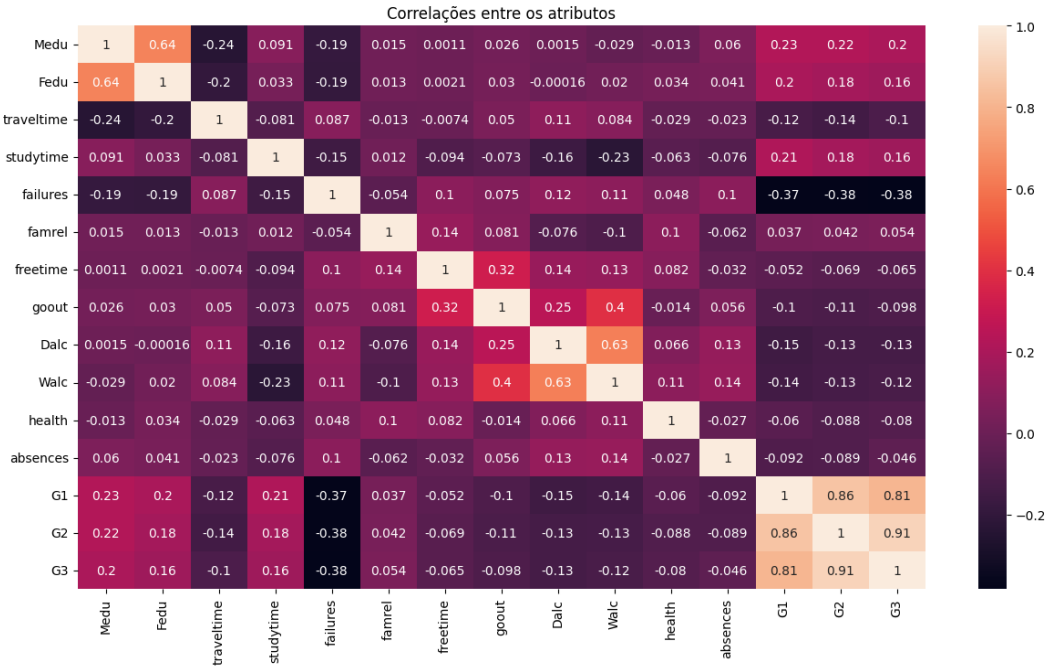
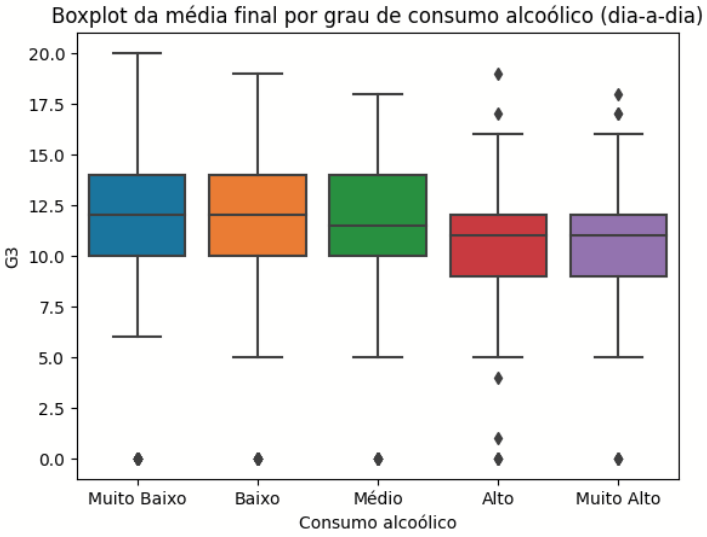


Figura 5: Correlação dos atributos no *dataset*

Percebe-se que este mapa de calor aponta diversas correlações entre os atributos, como, por exemplo, as notas G1, G2 e G3 são fortemente correlacionadas de maneira positiva entre si. Outros exemplos notáveis são o nível de educação materna/paterna ou o consumo alcoólico (ambos) e a frequência de passeios com amigos.

Consumo Alcoólico e Notas Finais



O consumo alcoólico diário (Dalc) possui uma correlação negativa de -0,12 em relação à nota final (G3), como ilustrado na Figura 5. Tal fato também pode ser observado no *boxplot* ilustrado pela Figura 6. Esta figura também demonstra que alunos com consumo alcoólico 'alto' e 'muito alto' possuem 3º quartil inferior aos demais.

Visando melhor demonstrar esta diferença, a Figura 7 demonstra distribuição normalizada de duas classes de alunos: (*i*) alunos consumo igual ou superior a 'médio'; (*ii*) alunos com o consumo inferior a médio.

Figura 6: Boxplot da nota final por nível de consumo

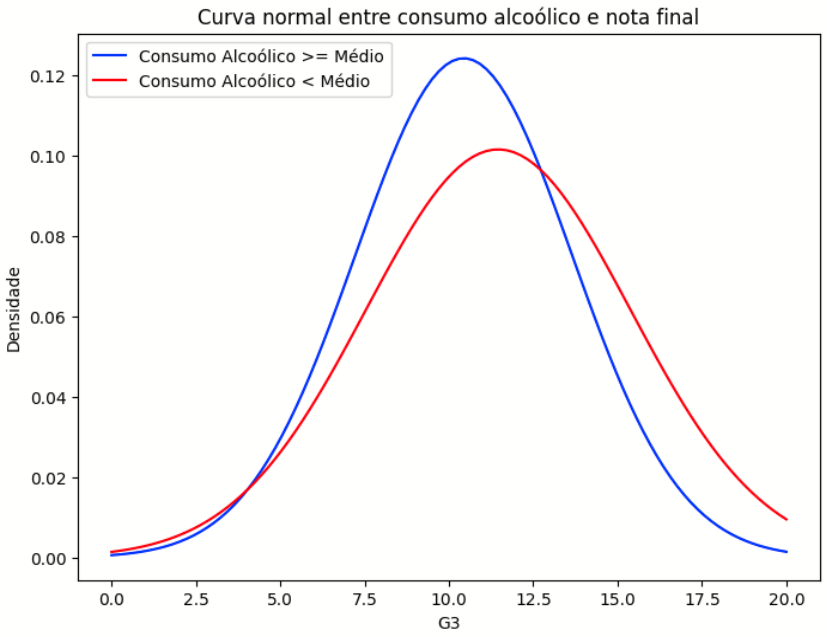


Figura 7: Curva normal dois grupos de consumo alcoólico

Testes de Hipóteses

Esta seção formaliza as hipóteses e verificar as mesmas utilizando testes estatísticos apropriados.

Hipótese

A partir da estipulação inicial descrita na introdução foi refinada e formalizada como:

- **H_0** (hipótese nula): não há uma correlação negativa entre o consumo alcoólico no dia-a-dia e a notas finais obtidas pelos indivíduos.
- **H_1** (hipótese alternativa): há uma correlação negativa significante entre o consumo alcoólico diário e notas finais obtidas.

É importante ressaltar que o nível de confiança (alfa) foi estabelecido como 5% na seção de Materiais e Métodos.

Verificação

Como previamente descrito na sub-seção pertinente a análise multi-variada, há uma correlação negativa de -0,12 entre o consumo alcoólico diário e notas finais. Esta correlação foi estabelecida por meio do método de Pearson. Ademais, o valor 'P' obtido neste teste foi de $2{,}65^{(10\text{^-}5)}$, isto é, 0,0000265, um valor inferior ao grau de significância previamente estabelecido. Sendo assim, há evidência suficiente para rejeitar a hipótese nula com base neste teste.

Entretanto, como um teste adicional, o método de comparação de variâncias ANOVA foi selecionado. De acordo com [10], este método é adequado para dados categóricos, como os níveis de consumo alcoólico. Em específico, foi utilizada a análise ANOVA uni-direcional (*one-way*), devido à variância de um único atributo (consumo alcoólico/Dalc) em relação a outro (nota final/G3). Tal teste resultou num valor 'F' de 6,24 e um valor 'P' de 0,0000579, isto é, também suportando a rejeição da hipótese nula, como ilustrado no Quadro 3. Por fim, é importante frisar que a métrica 'F', apesar de positiva, não contraria o coeficiente de correlação negativo obtido no teste de Pearson, uma vez que a métrica F não mensura a direção da correlação de variâncias.

```
grouped_data = df.groupby('Dalc')['G3'].apply(list)
# Perform the ANOVA test
f_statistic, p_value = stats.f_oneway(*grouped_data)
print('F-statistic:', f_statistic)
print('p-value:', p_value)
if p_value < 0.05:
    print("As diferências entre os grupos são estatisticamente
significantes.")
else:
    print("As diferências entre os grupos NÃO são estatisticamente
significantes.")
```

```
F-statistic: 6.24156863429061
p-value: 5.794348331246882e-05
As diferências entre os grupos são estatisticamente significantes.
```

Quadro 3: Resultado do teste ANOVA uni-direcional

Em resumo, há evidências estatísticas suficientes para rejeitar a hipótese nula (**H_0**) e afirmar que há uma correlação negativa entre o consumo alcoólico diário o desempenho acadêmico, contudo tal correlação não implica causalidade.

Conclusões

Este trabalho examinou a relação entre o consumo alcoólico e o desempenho acadêmico, e constatou uma correlação negativa fraca, porém estatisticamente significante, entre essas variáveis. A utilização de testes estatísticos, como a correlação de Pearson e a análise de variância (ANOVA), permitiu uma análise mais criteriosa e precisa dos dados.

A correlação negativa indica que quanto maior o consumo alcoólico, menor tende a ser o desempenho acadêmico dos estudantes. Contudo, é importante frisar que a correlação encontrada neste trabalho é fraca, e outros fatores como apoio familiar ou podem influenciar o desempenho acadêmico dos estudantes. Por fim, não é possível estabelecer um vínculo de causalidade entre estes atributos, apenas uma correlação.

Apesar disto, esses achados ressaltam a importância de se adotar uma abordagem consciente e responsável em relação ao consumo de álcool, especialmente durante a vida acadêmica. Além de políticas de conscientização em instituições de ensino, politicas públicas de restrição ao acesso desta substância podem reduzir os impactos negativos desta substância na população[8][9].

O autor sugere como um trabalho futuro o estudo das demais métricas contidas no *dataset* utilizado para a extração de novas correlações entre os atributos, assim como análise de novos conjuntos de dados mais detalhados e atualizados.

Referências Bibliográficas

[1] UCI MACHINE LEARNING (org.). Student Alcohol Consumption. 2016. Disponível em: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. Acesso em: 24 jun. 2023.

[2] CORTEZ, Paulo; SILVA, Alice Maria Gonçalves. Using data mining to predict secondary school student performance. 2008.

[3] NYU. NYU Study Examines Top High School Students' Stress and Coping Mechanisms. 2015. Disponível em: https://www.nyu.edu/about/news-publications/news/2015/august/nyu-study-examines-top-high-school-students-stress-and-coping-mechanisms.html. Acesso em: 25 jun. 2023.

[4] LEONARD, Noelle R. et al. A multi-method exploratory study of stress, coping, and substance use among high school youth in private schools. Frontiers in psychology, p. 1028, 2015.

[5] WIKIPEDIA (org.). Grading systems by country. Disponível em: https://en.wikipedia.org/wiki/Grading_systems_by_country. Acesso em: 28 jun. 2023.

[6] PORTUGAL. Decreto nº 50/2013, de 2013. Novo Regime de Disponibilização, Venda e Consumo de Bebidas Alcoólicas em Locais Públicos e em Locais Abertos Ao Público. Portugal, 16 abr. 2013. Disponível em: https://diariodarepublica.pt/dr/legislacao-consolidada/decreto-lei/2013-58361867. Acesso em: 27 jun. 2023.

[7] RITCHIE, Hannah; ROSER, Max. Alcohol Consumption. Elaborado por Our World in Data. Disponível em: https://ourworldindata.org/alcohol-consumption. Acesso em: 27 jun. 2023.

[8] SUDHINARASET, May; WIGGLESWORTH, Christina; TAKEUCHI, David T. Social and cultural contexts of alcohol use: Influences in a social–ecological framework. Alcohol research: current reviews, v. 38, n. 1, p. 35, 2016.

[9], Organização Mundial da Saúde. Fact Sheet: alcohol. Alcohol. 2022. Disponível em: https://www.who.int/news-room/fact-sheets/detail/alcohol. Acesso em: 27 jun. 2023.

[10] RODRIGUES, Leonardo. ANOVA: aprenda para o que serve, como calcular e em que momento utilizar essa variância!. Aprenda para o que serve, como calcular e em que momento utilizar essa variância!. 2019. Disponível em: https://www.voitto.com.br/blog/artigo/anova. Acesso em: 27 jun. 2023.

[11] MUKAKA, Mavuto M. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, v. 24, n. 3, p. 69-71, 2012.

Apêndice 1 - Sumário dos atributos do *dataset*

Nome	**Tipo**
'school' (escola)	String ● GP → Gabriel Pereira ● MS → Mousinho da Silveira
'sex' (sexo)	Binário (M/F)
'age' (idade)	Inteiro
'address' (endereço)	Binário (Urbano, Rural)
'famsize' (tamanho da família)	Binário (≥ 3 ou <3)
'Pstatus' (co-habitação com os pais)	Binária, 'T' (co-habitando) ou 'A' vivendo sozinho
'Medu' e 'Fedu (grau de educação materna/paterna)	Categórico (0-4): ● 0 → Nenhuma ● 1 → Até a 4ª série ● 2 → até a 9ª série ● 3 → Ensino Médio ● 4 → Ensino Superior
'Mjob' e 'Fjob' (tipo de emprego (materno/paterno)	Categórico: ● Professor(a), ● Funcionário na área de Saúde ● Funcionários em Serviços Civis (policia, administração) ● Nenhum (presente em casa) ● Outro
'reason' (motivo para escolher a escola)	Categórico: ● Proximidade ao lar ● Reputação da escola ● Curso ● Outra preferência
'guardian (guardão do aluno)'	Categórico: ● Mãe ● Pai ● Outro
'traveltime' (tempo de viagem)	Categórico ● 1 → Menos de 15 minutos ● 2 → entre 15 à 30 minutos ● 3 → entre 30 minutos a 1 hora ● 4 → superior a 1 hora
'studytime' (tempo alocado para estudo semanal)	Categórico ● 1 → Menos de 2 horas ● 2 → entre 2h a 5h ● 3 → entre 5h a 10h ● 4 → superior a 10 horas
'failures' (número de falhas em classes anteriores)	Categórico: ● 0 -> Menos que 3
'schoolsup' (recebe suporte da escola)	Binário (yes/no)

'famsup' (recebe suporte familiar)	Binário (yes/no)
'paid' (paga classes extras)	Binário (yes/no)
'activities' (atividades extra-curriculares)	Binário (yes/no)
'nursery' (teve jardim de infância)	Binário (yes/no)
'higher' (pretende cursar uma educação de nível superior)	Binário (yes/no)
'internet' (possui acesso a internet)	Binário (yes/no)
'romantic' (possui relacionamento romântico)	Binário (yes/no)
'famrel' (qualidade da relação com os parentes)	Categórico ● 1 → Péssimo ● 2 → Mau ● 3 → Média ● 4 → Boa ● 5 → Ótima
'freetime' (nível de tempo livre)	● 1 → Muito Baixo ● 2 → Baixo ● 3 → Médio ● 4 → Alto ● 5 → Muito Alto
'goout' (frequência com que sai com amigos)	Binário (yes/no)
'Dalc' (consumo alcoólico durante o dia-a-dia)	● 1 → Muito Baixo ● 2 → Baixo ● 3 → Médio ● 4 → Alto ● 5 → Muito Alto
'Walc' (consumo alcoólico durante finais de semana)	● 1 → Muito Baixo ● 2 → Baixo ● 3 → Médio ● 4 → Alto ● 5 → Muito Alto
'health' (grau de saúde)	Categórico ● 1 → Péssimo ● 2 → Mau ● 3 → Média ● 4 → Boa ● 5 → Ótima
'absences' (número de faltas)	Numérico 0-93
'G1' (nota parcial, primeiro período)	Numérico 0-20
'G2' (nota parcial, segundo período)	Numérico 0-20
'G3' (nota final)	Numérico 0-20