Análise do efeito de álcool no desempenho acadêmico

1st Raul Lomonte Figueiredo
FGV EMAp
Rio de Janeiro, Brazil
raullomonte13@gmail.com

***Resumo*—Este estudo busca elucidar a associação entre o consumo de álcool em adolescentes e o seu desempenho acadêmico, uma questão social e educacional de grande importância com implicações potencialmente significativas para o desenvolvimento do país. A educação básica, reconhecida como um dos pilares fundamentais para o crescimento e desenvolvimento de qualquer nação, torna-se o foco desta investigação. É vital identificar e compreender os fatores que podem influenciar o desempenho acadêmico dos estudantes, para assim promover estratégias efetivas de aprendizado e melhorar a qualidade da educação.**

Por outro lado, o consumo de álcool na adolescência tem sido objeto de numerosos estudos devido à sua prevalência e aos efeitos potencialmente prejudiciais. De acordo com dados fornecidos pelo Centro de Informações sobre Saúde e Álcool (CISA), estima-se que, no Brasil, cerca de 63,3% dos estudantes entre 13 e 17 anos já experimentaram bebida alcoólica, com 55,9% dos estudantes de 13 a 15 anos reportando experimentação. Este estudo argumenta que essa prevalência considerável tem o potencial de impactar o desempenho acadêmico dos alunos e, portanto, merece ser investigada.

Para realizar tal análise, este trabalho propõe o uso de métodos estatísticos para explorar a correlação entre o consumo de álcool e o desempenho dos alunos, especificamente em alunos do ensino secundário, normalmente na faixa etária de 15 a 18 anos. O objetivo é identificar a existência de um efeito claro do consumo de álcool na performance acadêmica.

Neste contexto, foi utilizada uma base de dados que contém informações detalhadas sobre os alunos de duas escolas em Portugal, incluindo os hábitos de consumo de álcool e os resultados acadêmicos, com o intuito de entender melhor a situação. A análise foi realizada empregando o modelo de regressão logistica, um método estatístico popular em pesquisas que lidam com variáveis dependentes categóricas - neste caso, os alunos foram classificados como 'aprovados' ou 'reprovados'.

No entanto, os resultados obtidos na análise não permitiram confirmar de forma convincente um efeito relevante do consumo de álcool no desempenho acadêmico. A principal razão para essa falta de clareza reside na alta correlação encontrada entre as variáveis analisadas. Por exemplo, descobrimos que o consumo de álcool estava significativamente relacionado com outros fatores como o sexo e o tempo dedicado ao estudo. Essa forte interdependência entre as variáveis tornou difícil isolar o efeito individual do consumo de álcool no desempenho acadêmico. Dada a natureza complexa dessas relações, torna-se necessária uma análise mais sofisticada e específica para discernir com precisão o impacto do consumo de álcool na performance escolar. Estes resultados realçam a complexidade do tema e a necessidade de abordagens de pesquisa mais refinadas para desvendar a intrincada rede de fatores que influenciam o desempenho acadêmico dos adolescentes.

I. Introdução

A educação, ao longo do tempo, tem sido reconhecida como um dos pilares fundamentais para o desenvolvimento econômico e social de qualquer nação. Na busca contínua pela excelência educacional, reside o desafio constante de melhorar a qualidade do ensino, principalmente nas escolas de ensino fundamental e médio. A questão que se coloca não é apenas como transmitir conhecimento de forma eficaz, mas também como inspirar os alunos a buscar ativamente o conhecimento por conta própria. Esta questão é especialmente complexa quando consideramos a influência de fatores externos ao ambiente escolar, que podem afetar a motivação, o engajamento e o desempenho dos alunos.

No entanto, essa análise da educação não estaria completa sem considerar questões de saúde pública que afetam diretamente a população jovem. Neste contexto, o consumo de álcool entre os jovens é um tema de crescente preocupação, tanto para a comunidade educacional quanto para os profissionais de saúde. É de conhecimento amplo que o consumo de bebidas alcoólicas pode ter efeitos nocivos substanciais sobre o organismo em desenvolvimento do adolescente. De acordo com o Centro de Informações sobre Saúde e Álcool (CISA) [1], no Brasil, cerca de 63,3% dos estudantes entre 13 e 17 anos já experimentaram álcool, e entre os estudantes de 13 a 15 anos, o índice é de 55,9%. Estes dados ilustram a extensão da exposição ao álcool nesta faixa etária e sugerem um potencial impacto significativo no desempenho acadêmico dos alunos, um aspecto que precisa ser cuidadosamente investigado.

Neste trabalho, nos propomos a investigar essa relação, focando especificamente na influência direta do consumo de álcool no rendimento escolar dos alunos, medido através da sua aprovação ou reprovação no ano letivo. Para realizar este estudo, nos baseamos em um conjunto de dados robusto, contendo informações sobre alunos do ensino secundário (o equivalente ao ensino médio no Brasil) de duas escolas portuguesas. Estes dados nos fornecem uma visão detalhada dos aspectos sociais e econômicos dos estudantes, bem como do seu desempenho acadêmico em três avaliações durante o ano, nas disciplinas de Português e Matemática.

Para efeitos desta análise, nos concentramos na terceira nota de cada aluno, dado que é a avaliação final, logo, se esta nota fosse igual ou superior a 10, o aluno seria considerado aprovado, e caso contrário, reprovado. Com base

neste critério, construímos e avaliamos diversos modelos de regressão logística, visando identificar as variáveis que mais contribuem para o desempenho do aluno. O modelo que melhor se ajustou aos dados alcançou uma acuracia de 89,6% na predição de aprovação de português e 79.8% em matemática, utilizando apenas oito variáveis.

Aprofundando ainda mais a nossa investigação, focamos particularmente nas variáveis do consumo de álcool e no seu efeito sobre o rendimento escolar. Dada a prevalência do consumo de álcool entre os estudantes e o seu potencial impacto na aprendizagem, acreditamos que este é um aspecto crítico que precisa ser abordado. Através desta investigação, inicialmente esperavamos lançar uma luz sobre como o consumo de álcool pode afetar o desempenho acadêmico dos alunos e fornecer insights para políticas de saúde pública e estratégias educacionais.

Este artigo fundamenta-se significativamente no trabalho do conceituado professor português Paulo Cortes. Cortes é um acadêmico, com inúmeros artigos publicados, e tem dedicado muito de seu tempo à disponibilização de diversos conjuntos de dados coletados, inclusive o que esta sendo usados nesta pesquisa. Ele é também o autor da biblioteca em R, conhecida como rminer, e tem alguns livros publicados em seu nome.

O artigo de referência para este estudo é intitulado "Using data mining to predict secondary school student performance" [2], que foi publicado em 2008 e, até à data, acumulou mais de 800 citações. Neste trabalho, Cortes emprega o mesmo conjunto de dados que estamos usando, mas sua abordagem é muito diferente. Seu foco reside na criação de um modelo de machine learning utilizando várias técnicas, incluindo Decision Trees, Random Forest, Neural Networks e Support Vector Machines. Ele ajusta estes modelos de três maneiras distintas: para classificação binária (aprovado/reprovado), classificação em cinco níveis de desempenho (onde 1 é ótimo e 5 é insuficiente) e através de uma abordagem de regressão.

A diferença de abordagem entre o trabalho de Cortes e o presente estudo é notavelmente evidente. Enquanto Cortes opta por uma abordagem de machine learning, caracterizada por alimentar os dados através de um mecanismo de "caixa preta", este trabalho adota uma perspectiva diferente. Nosso objetivo é explorar a relação entre determinados eventos externos e o desempenho na avaliação, buscando entender as influências subjacentes que podem afetar os resultados do aluno. Esta pesquisa, portanto, procura abrir essa "caixa preta", tentando desvendar as complexidades que envolvem o desempenho dos estudantes.

II. Metodologia

A. Base de dados

Em Portugal, o ensino secundário tem a duração de 3 anos, precedidos por 9 anos de ensino básico, e é seguido pelo ensino superior, semelhante ao ensino médio brasileiro. Durante o ensino secundário, a cada ano letivo, os estudantes são avaliados em três períodos de tempo, sendo a última avaliação (denominada G3 no banco de dados) correspondente à nota final. As notas das avaliações variam de 0 a 20, onde 20 é a nota máxima e 0 é a nota mais baixa. Caso a nota final seja menor que 10, o aluno é reprovado; caso contrário, é aprovado.

Os dados utilizados neste trabalho foram coletados por Paulo Cortes durante o ano escolar de 2005-2006, em duas escolas públicas da região de Alentejo, em Portugal. O conjunto de dados foi construído a partir de duas fontes: boletins escolares, que contêm as notas das três avaliações e o número de faltas, e questionários, utilizados para complementar as informações anteriores. Foram consideradas 29 perguntas, todas elas fechadas, relacionadas a variáveis demográficas (por exemplo, educação da mãe), sociais (por exemplo, consumo de álcool) e relacionadas à escola (por exemplo, número de reprovações em anos anteriores). Os dados das pesquisas foram integrados aos boletins das disciplinas de Matemática (com 395 exemplos) e Língua Portuguesa (com 649 exemplos).

O conjunto de dados total contem 33 variaveis, esses são seus nomes e suas definições:

1) **school** - escola do aluno (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
2) **sex** - sexo do aluno (binário: 'F' - feminino ou 'M' - masculino)
3) **age** - idade do aluno (numérico: de 15 a 22)
4) **address** - tipo de endereço residencial do aluno (binário: 'U' - urbano ou 'R' - rural)
5) **famsize** - tamanho da família (binário: 'LE3' - menor ou igual a 3 ou 'GT3' - maior que 3)
6) **Pstatus** - status de convivência dos pais (binário: 'T' - vivendo juntos ou 'A' - separados)
7) **Medu** - educação da mãe (numérico: 0 - nenhum, 1 - educação primária (4ª série), 2 - 5ª a 9ª série, 3 - educação secundária ou 4 - educação superior)
8) **Fedu** - educação do pai (numérico: 0 - nenhum, 1 - educação primária (4ª série), 2 - 5ª a 9ª série, 3 - educação secundária ou 4 - educação superior)
9) **Mjob** - profissão da mãe (nominal: 'teacher', cuidados 'health' relacionados, 'services' civis (por exemplo, administrativos ou policiais), 'at_home' ou 'other')
10) **Fjob** - profissão do pai (nominal: 'teacher', cuidados 'health' relacionados, 'services' civis (por exemplo, administrativos ou policiais), 'at_home' ou 'other')
11) **reason** - motivo para escolher essa escola (nominal: 'close to home', 'reputation' da escola, preferência de 'course' ou 'other')
12) **guardian** - responsável pelo aluno (nominal: 'mother', 'father' ou 'other')
13) **traveltime** - tempo de deslocamento de casa para a escola (numérico: 1 - $<$15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - $>$1 hora)
14) **studytime** - tempo semanal de estudo (numérico: 1 - $<$2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, ou 4 - $>$10 horas)
15) **failures** - número de reprovações em anos anteriores (numérico: n se 1 $<=$ n $<$ 3, senão 4)
16) **schoolsup** - apoio educacional adicional (binário: sim ou não)

17) **famsup** - apoio educacional familiar (binário: sim ou não)
18) **paid** - aulas extras pagas dentro da disciplina do curso (Matemática ou Português) (binário: sim ou não)
19) **activities** - atividades extracurriculares (binário: sim ou não)
20) **nursery** - frequentou a creche/escola infantil (binário: sim ou não)
21) **higher** - deseja cursar ensino superior (binário: sim ou não)
22) **internet** - acesso à Internet em casa (binário: sim ou não)
23) **romantic** - em um relacionamento romântico (binário: sim ou não)
24) **famrel** - qualidade dos relacionamentos familiares (numérico: de 1 - muito ruim a 5 - excelente)
25) **freetime** - tempo livre após a escola (numérico: de 1 - muito baixo a 5 - muito alto)
26) **goout** - sair com amigos (numérico: de 1 - muito baixo a 5 - muito alto)
27) **Dalc** - consumo de álcool durante a semana (numérico: de 1 - muito baixo a 5 - muito alto)
28) **Walc** - consumo de álcool nos fins de semana (numérico: de 1 - muito baixo a 5 - muito alto)
29) **health** - estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom)
30) **absences** - número de faltas na escola (numérico: de 0 a 93)
31) **G1** - nota do primeiro período (numérico: de 0 a 20)
32) **G2** - nota do segundo período (numérico: de 0 a 20)
33) **G3** - nota final (numérico: de 0 a 20)

E aqui, as variaveis que vamos focar mais serão Dalc e Walc.

B. Modelos, ajuste e avaliação

Neste trabalho, o foco está na análise da correlação entre o consumo de álcool e o desempenho escolar dos alunos, especificamente na aprovação ou reprovação. No processo de escolha do modelo analítico, optamos pela regressão logística, dado que não encontramos razões substanciais que nos impedissem de utilizar esta técnica. Destaca-se que a regressão logística, devido à sua eficácia na classificação binária, correspondeu perfeitamente às nossas necessidades, sem introduzir complexidade excessiva ao modelo. Portanto, esta emergiu como a opção mais adequada para a nossa pesquisa.

No início do estudo, considerei a abordagem das duas outras técnicas discutidas no artigo de Paulo Cortes: a regressão linear e a classificação multiclasse. A ideia inicial de aplicar a regressão linear era muito boa, tinha como base lógica a tentativa de observar o coeficiente angular do consumo de álcool para compreender a relação direta entre o consumo de álcool e as notas escolares. No entanto, apesar de várias tentativas, não conseguimos ajustar um modelo satisfatório, já que todos apresentavam um erro padrão e MSE consideravelmente altos. Este fator inviabilizou a consideração desses modelos para a extração de conclusões significativas, ao contrário da regressão logística, que produziu resultados excelentes.

Quanto à classificação multiclasse, não acreditamos que ela agregaria valor significativo ao estudo. As classes propostas por Cortes, que vão de notas 20 a 16 para a primeira classe, 15 a 14 para a segunda, 13 a 12 para a terceira, 11 a 10 para a quarta, e 9 a 0 para a quinta, parecem demasiado abstratas. Além disso, essa abordagem pareceu estar muito próxima da regressão, porém ”desconsiderando”as pontuações nas extremidades. Assim, optamos por utilizar apenas a regressão logística para prever se um aluno seria aprovado ou reprovado, o que se mostrou mais apropriado para este trabalho.

Em relação ao ajuste de modelos, todos os que foram testados utilizaram ajustes baseados em máxima verossimilhança, mais especificamente, empregando a função 'glm' do pacote base do R. Esta função demonstrou-se extremamente eficiente, satisfazendo todas as necessidades requeridas para a realização deste trabalho, considerando o modelo escolhido.

Incialmente, cogitei utilizar uma abordagem mais bayesiana, por meio da função 'stan_glm'. Contudo, enfrentei uma série de problemas ao tentar instalar essa biblioteca no RStudio do meu sistema Linux, mesmo tendo conseguido instalar essa mesma biblioteca anteriormente no Windows sem dificuldades. Esse contratempos consumiram um tempo considerável.

Diante deste desafio, voltei minha atenção para a abordagem frequentista. Ao avaliá-la mais profundamente, percebi que ela poderia, de fato, atender aos objetivos do meu trabalho. Assim, a abordagem frequentista emergiu como a melhor opção e caminho a ser seguido para este estudo.

Para a realização da avaliação dos modelos, foram consideradas duas abordagens: a bondade de ajuste usando o AICc e a capacidade preditiva utilizando a acurácia com LOO (leave one out). A utilização do AICc como critério de avaliação é uma abordagem interessante, pois leva em consideração a complexidade do modelo, penalizando modelos desnecessariamente mais complexos que não agregam informação suficiente para justificar o aumento na complexidade. Dessa forma, é possível desenvolver um modelo de qualidade sem que ele se torne absurdamente complexo. Além disso, como o conjunto de dados não é muito grande, restringir a complexidade do modelo também ajuda a evitar overfitting.

A decisão de estimar a capacidade preditiva utilizando a acurácia com LOO foi tomada visando aproveitar ao máximo os dados disponíveis. Considerando que o conjunto de dados é composto por 649 linhas em um caso e 395 linhas em outro, não seria tão interessante dividir os dados em apenas um conjunto de treino e teste. Durante as aulas, foi discutido o poder do LOO, e essa abordagem foi considerada interessante e enriquecedora para o trabalho, mesmo que possa demandar mais recursos computacionais.

III. Resultados

A. Analise exploratoria

Após uma análise abrangente da distribuição dos dados em seus respectivos domínios, comecei a concentrar-me nas duas variáveis de interesse: 'Dalc' e 'Walc'. Observou-se

inicialmente que a distribuição dos alunos em cada categoria de intensidade de consumo de álcool é altamente assimétrica, como ilustrado na Figura 1. Isso é bastante compreensível, uma vez que estamos lidando principalmente com menores de idade, o que significa que é de se esperar que haja poucos alunos que consumam álcool de forma intensa. Com base nesse resultado, prosseguimos com a análise, sempre utilizando porcentagens em vez de valores absolutos, a fim de evitar impressões falsas sobre a relação do consumo de álcool. É importante observar também que a escassez de dados sobre os alunos que consomem álcool de forma muito intensa é um desafio, com apenas 17 alunos na categoria 4 e 17 alunos na categoria 5 de consumo de álcool durante a semana. Isso dificulta a compreensão de como o efeito se manifesta entre os bebedores. Seria interessante ter uma amostra com mais dados nesse aspecto. As imagens e números mencionados durante essa parte referem-se aos dados de português, uma vez que há uma quantidade maior de dados disponíveis, e quase todos os alunos presentes nos dados de matemática também estão presentes aqui, foram feitas as mesmas analises no outro conjunto de dados e os resultados foram extremamente proximos.

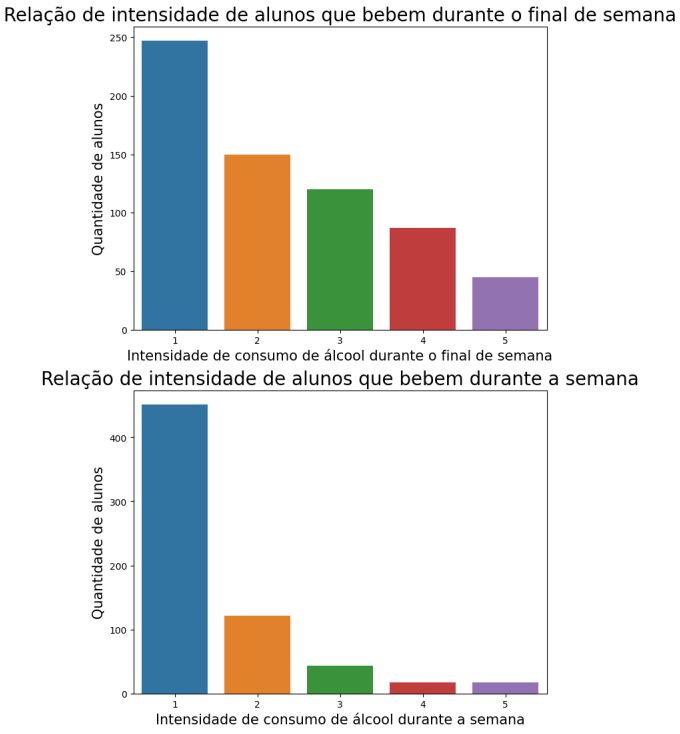

Figura 1. Representação visual da quantidade de alunos que consomem bebidas durante e nos finais de semana e sua intensidade.

Um próximo passo importante foi buscar compreender a correlação entre as duas variáveis de interesse. É possível observar na Figura 2 uma relação bastante significativa, o que representou um desafio, dificultando a análise do efeito de cada variável individualmente. Unificar as duas variáveis em uma única medida aumentaria consideravelmente a complexidade do trabalho. No entanto, ciente dessa dificuldade, prossegui com a exploração a fim de visualizar como ambas as variáveis se relacionam com as notas.

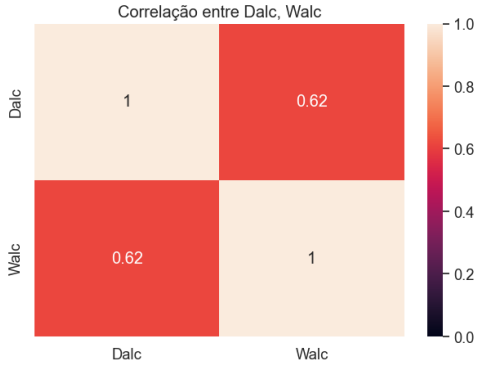

Figura 2. Correlação entre o consumo de álcool durante e aos finais de semana.

Mais um longo processo de exploração foi realizado até chegar na imagem que considerei mais interessante e que mais me incentivou a interromper a exploração e começar a desenvolver os modelos. A imagem em questão ilustra a relação percentual entre alunos aprovados, divididos em cada categoria de intensidade de consumo de álcool. Ao observar o gráfico representado na Figura 3, fiquei extremamente confiante de que conseguiria identificar um efeito claro do consumo de álcool em relação à aprovação, ou até mesmo à nota direta do aluno. Portanto, decidi interromper minha análise exploratória e direcionar meus esforços para a modelagem. Toda essa análise exploratória foi realizada em Python, utilizando as bibliotecas pandas, matplotlib e seaborn. Já a etapa de modelagem, que será apresentada a seguir, foi feita em R.

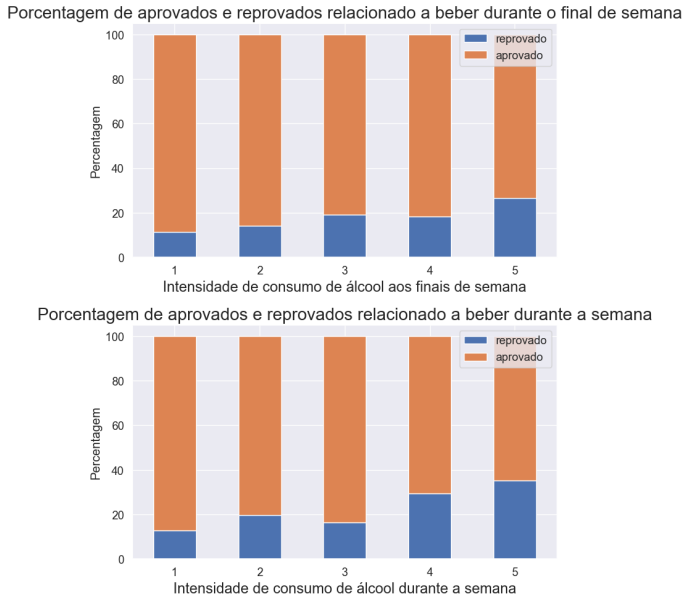

Figura 3. Porcentagem de aprovados separada por intensidade de consumo de álcool.

B. Ajuste de modelos

Ao iniciar o processo de ajuste dos modelos, precisei tomar uma série de decisões críticas enquanto me familiarizava com os dados por meio do ambiente de programação R. Inicialmente, os dados foram normalizados, uma abordagem inspirada no tratamento inicial de dados ilustrado no capítulo 12.7 do livro 'Regression and Other Stories'. Tal normalização revelou-se fundamental, visto que, embora algumas variáveis parecessem ter um valor insignificante à primeira vista, após normalização, seus verdadeiros valores tornaram-se evidentes. Essa mudança deve-se à escala na qual as variáveis estavam originalmente presentes. A normalização foi realizada utilizando a função "scale"do R, aplicada às variáveis necessárias.

Com os dados já devidamente normalizados, iniciei minha jornada em busca dos melhores modelos. Utilizando a função "glm"do R base, baseada no ajuste em máxima verossimilhança, defini como variável alvo uma que não vem por padrão na base de dados, denominada "aprovado". Esta está diretamente ligada à variável G3; quando G3 é menor do que 10, "aprovado"tem valor 0 e, caso contrário, tem valor 1. No parâmetro da família, utilizei "binomial(link = "logit")", pois meu objetivo é realizar uma regressão logística. Os dados passados como parâmetros são aqueles já contendo a variável "aprovado"e devidamente normalizados.

Em uma das etapas exploratórias, foi gerada uma extensa tabela de correlações entre variáveis. Utilizei esta como um mapa para guiar a escolha das variáveis em diferentes modelos. Os modelos foram avaliados de duas formas: uma analisando a bondade do ajuste e a outra, a precisão da predição do modelo, através do AICc e da Acurácia no LOO (Leave-One-Out). Dessa maneira, pudemos avançar em direção a um modelo mais preciso, sem incorrer em uma complexidade excessiva.

Conforme sugerido pelo professor, as avaliações de português e matemática foram tratadas separadamente. Os ajustes e buscas foram primeiramente testados no conjunto de dados da avaliação de português, seguidos pelo conjunto de matemática. A razão para essa abordagem deve-se ao fato de que quase todos os alunos catalogados em matemática estão inclusos no conjunto de dados de português - mas a recíproca não é verdadeira. Além disso, o conjunto de português possui um volume maior de dados, proporcionando maior margem para testes e análises de resultados.

Após a avaliação de um número impressionante de modelos, o modelo mais eficiente, considerando as variáveis do consumo de álcool, foi:

mod $\leftarrow$ glm(aprovado $\sim$ failures + higher + school + studytime +age + absences + Dalc + Walc , family = binomial(link = "logit"), data = datastd_aprovado)

O modelo ajustado leva em conta a quantidade de reprovações passadas do aluno, se ele deseja cursar o ensino superior, a escola em que ele está, o tempo de estudo, a idade, número de faltas, a quantidade de álcool que ele consome durante a semana e nos finais de semana. Mesmo tendo algumas variáveis que, ao analisar o seu coeficiente, não parecem muito significativas, ao alterar qualquer uma dessas variáveis do modelo, ele perde em AICc e/ou em acurácia. A acurácia final do modelo foi de 89.6% no conjunto de dados considerando a prova de português e 79.8% no conjunto de matemática. O AICc foi de 426.80 para o conjunto de treino de português e 467.75 para o de matemática. A fim de visualizar a qualidade do modelo, gerei a imagem da curva ROC dele. A curva ROC não foi um critério para escolha do modelo dentre os anteriores, mas foi um bom teste de sanidade para ver que nosso modelo realmente faz sentido e faz o que é proposto. Podemos ver o resultado na figura 4.

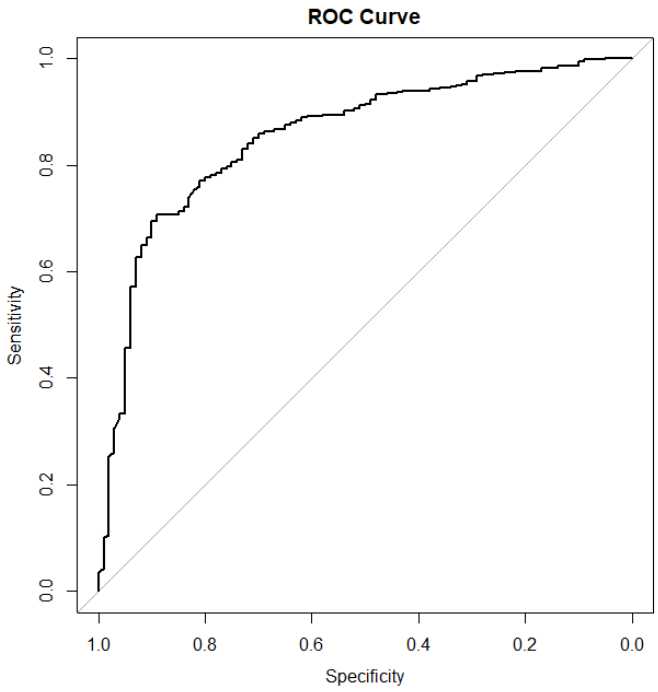

Figura 4. Curva ROC do modelo proposto.

O que mostra um resultado interessante, dado a missão do modelo de classificar se o aluno irá passar ou não na prova, com poucas informações, é uma missão bem difícil e o modelo está conseguindo se sair bem. Porém, a felicidade acabou logo quando fui de fato começar a analisar o efeito do álcool na aprovação ou reprovação do aluno. A figura 5 me desapontou um pouco, dado que, como mostrado na análise exploratória, havia uma grande chance de existir relação entre o consumo de álcool e a reprovação do aluno. Ver esse resultado me fez parar e refletir sobre todo o trabalho, conversar bastante com o monitor para entender de forma correta esse resultado contraditório e assim voltar ao objetivo de buscar o feito após entender o motivo desse resultado.

Depois de muita discussão, chegamos à conclusão de que isso poderia ser causado pela dependência entre as variáveis. Já era de conhecimento que Dalc e Walc estavam relacionadas, então foram criados modelos com apenas uma delas e mesmo assim não parecia ter efeito. Isso me fez voltar à análise exploratória e tentar buscar mais relações diretas entre as variáveis escolhidas para o modelo. Consegui então enxergar o problema mais a fundo. Encontrei exemplos claros, como na figura 6 e figuras 7, de casos onde as variáveis de consumo

```
> summary(mod_geral)

Call:
glm(formula = aprovado ~ failures + higher + school + studytime +
    age + absences + Dalc + Walc, family = binomial(link = "logit"),
    data = datastd_aprovado)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8579   0.2115   0.2873   0.5196   1.9836

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.29205    0.16657  13.761  < 2e-16 ***
failures    -0.67276    0.11318  -5.944 2.78e-09 ***
higher       0.40139    0.10325   3.887 0.000101 ***
school      -0.88327    0.13587  -6.501 7.98e-11 ***
studytime    0.19218    0.14351   1.339 0.180542
age          0.26576    0.14508   1.832 0.066988 .
absences    -0.28568    0.12356  -2.312 0.020776 *
Dalc        -0.05505    0.14710  -0.374 0.708247
Walc        -0.18159    0.15771  -1.151 0.249548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 557.79  on 648  degrees of freedom
Residual deviance: 408.52  on 640  degrees of freedom
AIC: 426.52
```

Figura 5. Output gerado pela função summary do modelo proposto.

de álcool estão diretamente ligadas às outras variáveis, como, por exemplo, essa relação clara entre alunos que não querem ingressar no ensino superior e a quantidade de álcool que bebem.

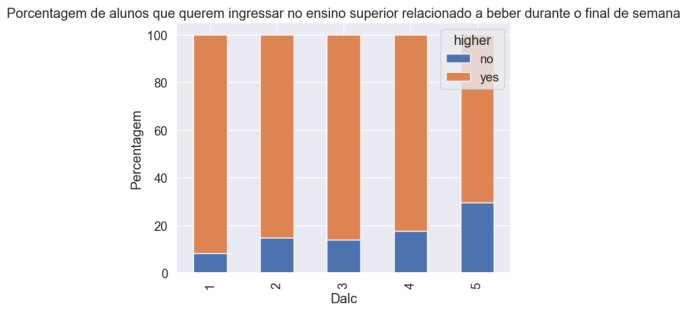

Figura 6. Relação entre beber aos finais de semana e o desejo de ingressar no ensino superior.

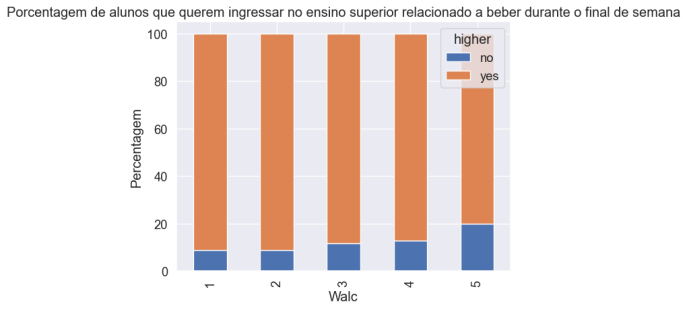

Figura 7. Relação entre beber durante a semana e e o desejo de ingressar no ensino superior..

Essa relação intrínseca entre as diversas variáveis dificultou a análise desejada inicialmente e que à primeira vista tinha enorme potencial. Foram cogitadas algumas alternativas de isolar o efeito da variável das outras variáveis relacionadas, mas nada viável para essa entrega, dado meu conhecimento prévio e a data limite de entrega. Portanto, não foi possível fazer a real análise do efeito do álcool no desempenho escolar. Infelizmente, encontrei diversos problemas e mesmo contornando vários, alguns foram impeditivos para a conclusão do objetivo inicial.

IV. Conclusão

Em conclusão, este estudo empreendeu um profundo mergulho na influência do consumo de álcool na aprovação ou reprovação dos alunos do ensino secundário, correspondente a faixa etaria do ensino médio brasileiro. Ao longo do trabalho, exploramos diversas técnicas de modelagem estatística, incluindo o uso de um modelo de regressão logística baseado em abordagem frequentistas para ajustar o modelo, além de estudar os outros modelos e entender seus prós e contras. Foi necessaria a utilização do Critério de Informação de Akaike Corrigido (AICc) para uma boa avaliação do modelo, buscando limitar a complexidade e tentando melhorar o quanto o modelo explica os dados. E além disso a validação cruzada Leave-One-Out (LOO) para avaliar os modelos de uma forma robusta e uma solução para nosso conjunto de dados não tão grande com muitas features, previnindo um possivel overffiting.

A riqueza dos dados coletados proporcionou uma grande oportunidade para uma análise exploratória extensiva, com uma quantidade expressiva de informações de cada aluno, o que deixa toda a exploração mais interessante e rica. Investiguei as relações entre variáveis diversas, tentando entender melhor como fatores externos podem influenciar tanto no desempenho do aluno, e mostrando como fica claro que a educação do aluno está longe de depender apenas da escola, mas sim do conjunto completo de vivências do aluno. Trabalhei intensamente no ajuste dos coeficientes do nosso modelo até chegarmos ao produto final: um modelo de previsão robusto, capaz de prever a aprovação ou reprovação de um aluno com informações limitadas e com uma acurácia interessante.

No entanto, o objetivo principal, que era identificar o impacto do consumo de álcool na aprovação dos alunos, não foi tão bem-sucedido quanto gostaríamos. A análise dos dados revelou que as variáveis, de maneira geral, estavam inter-relacionadas, o que reflete a complexidade inerente do mundo real. Foram discutidas várias estratégias com o monitor da disciplina para abordar essa interdependência, incluindo a possível utilização de Análise de Componentes Principais (PCA) e Modelagem de Equações Estruturais (SEM). No entanto, o tempo e a complexidade adicional necessários para implementar essas técnicas não permitiram sua aplicação dentro do prazo do projeto.

Assim, embora este trabalho termine com um tom um tanto melancólico, devido à nossa incapacidade de determinar definitivamente o efeito das variáveis relacionadas ao álcool, e do álcool de forma geral, acredito que o trabalho foi muito proveitoso para meu conhecimento de estatistica de forma bem geral. Acredito que como ponto principal do trabalho, aprendi muito ao longo deste projeto, tanto sobre as limitações de

diferentes técnicas e abordagens, tanto na escolha de modelos quanto na avaliação dos mesmos. Compreendi o quão desafiador pode ser realizar uma exploração adequada dos dados e o quanto isso influencia o futuro das pesquisas. Entender completamente os dados com os quais se lida, mesmo sem ser um especialista na área, é um desafio significativo. No entanto, deixo em aberto a questão inicial do trabalho, esperando que futuros pesquisadores se aventurem a empregar abordagens mais sofisticadas, como PCA e SEM, para desvendar a intrincada teia de interdependências entre as variáveis.

O desafio de entender o efeito do álcool na aprovação dos alunos está longe de ser resolvido e este trabalho, mesmo com suas limitações, adicionou mais uma peça ao complexo quebra-cabeça da educação e da vida dos adolescentes. Ficou claro que é necessário uma análise mais aprofundada e talvez um projeto mais amplo, com prazo suficiente para aplicar técnicas mais avançadas, para chegar a conclusões mais definitivas sobre o impacto do consumo de álcool na educação dos alunos.

Em suma, embora não tenha alcançado nossas metas originais, acredito que esse estilo de trabalho aberto é muito interessante para entender realmente como funciona a modelagem estatística na prática. Para desenvolver todo esse trabalho e compreender o que estava fazendo, foi necessário revisar diversos conteúdos passados em aula e estudar outros que não foram abordados, a fim de compreender melhor o funcionamento das técnicas ou entender o motivo por trás de cada escolha feita no trabalho. Durante o projeto, além do contato inicial para esclarecer dúvidas com o professor, tive diversas conversas com o monitor para tentar abordar essas limitações. No entanto, a complexidade do trabalho foi aumentando gradualmente até ultrapassar minha capacidade de resolver o problema original dentro do prazo de entrega. Os próximos passos para o trabalho seriam a aplicação do PCA, como sugerido pelo monitor, para tentar isolar ainda mais o efeito das variaveis Dalc e Walc, e o ápice do trabalho seria a implementação do SEM para conseguir definir a variável de consumo de álcool, combinando Dalc e Walc, e assim obter a resposta real para a pergunta inicial: ”Qual o efeito do álcool no desempenho dos alunos?”

REFERÊNCIAS

[1] Centro de Informações sobre Saúde e Álcool (CISA) - https://cisa.org.br/

[2] Cortez, Paulo, and Alice Maria Gonçalves Silva. ”Using data mining to predict secondary school student performance.”(2008).

[3] Regression and Other Stories - Andrew Gelman, Jennifer Hill, Aki Vehtari