Projeto Prático de Aprendizagem de Máquina

Adriano Nascimento da Paixão, 20180015150

Carlos Alberto Alves de Meneses, 201800032

2022-11-26

**Efeitos do Álcool no Estudo**

**1 Introdução**

Todos sabemos que o consumo em excesso de bebidas alcólicas provoca reaçòes indesejáveis no nosso organismo.

Alguns copos a mais e você sente dificuldade para falar com desenvoltura e de coordenação e equilíbrio, e para piorar, sua capacidade de julgamento já não é tão confiável.

Segundo Galvao (2017), O excesso de álcool no cérebro leva a efeitos psíquicos como redução da concentração, da atenção, da memória recente e da capacidade de julgamento.

Para Mônica (2018), o álcool é um depressor do sistema nervoso central, ou seja, uma substância que diminui a atividade do cérebro, alterando a ação de neurotransmissores, como o ácido gama-aminobutírico, o glutamato e a serotonina. Conforme a pessoa ingere a bebida, o organismo reage de uma determinada forma, seguindo alguns estágios.

Quando a concentração de álcool no sangue é baixa (entre 0,01 e 0,12 gramas/100 mililitros), o indivíduo tende a ficar desinibido, relaxado e eufórico. À medida que essa quantidade aumenta, outras reações aparecem, como lentidão dos reflexos, problemas de atenção, perda de memória, alterações na capacidade de raciocínio e falta de equilíbrio. Em níveis muito altos (a partir de 0,40 gramas/100 mililitros), pode haver intoxicação severa e parada cardiorrespiratória, com possibilidade de sequelas neurológicas e até mesmo morte.

Além desses efeitos visíveis e imediatos, o consumo exagerado de álcool, principalmente ná infância e adolescência, pode prejudicar o desenvolvimento cerebral, inibir o crescimento de novos neurônios e causar lesões permanentes, além de ser um fator de risco para a depressão ou outro transtorno mental.

**2 Objetivo**

O objetivo é criar um modelo preditivo para responder a seguinte pergunta: “Como séria o desempenho das notas dos alunos nas disciplinas de Português e Matemática em relação ao consumo do álcool?”

**3 Método**

Foram análisados bancos de dados sobre o tema proposto (*Alcohol Effects On Study*), disponíveis no site <https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study>. Utilizou-se o software R para leitura e análise descritiva dos dados.

O que na realidade queremos saber é: ” É possível usar características quantitativas para predizer as notas dos alunos na disciplina de português e matemática (G3) face ao concumo de álcool?”

**4 Predição para as notas de português**

**4.1 Amostra de Entrada**

A tabela a seguir, nos mostra as primeiras seis linhas do banco de dados *Portuguese.csv*.

**5 Análise Descritiva**

## # A tibble: 649 × 33  
## school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason  
## <chr> <chr> <dbl> <chr> <chr> <chr> <dbl> <dbl> <chr> <chr> <chr>   
## 1 GP F 18 U GT3 A 4 4 at\_home teach… course  
## 2 GP F 17 U GT3 T 1 1 at\_home other course  
## 3 GP F 15 U LE3 T 1 1 at\_home other other   
## 4 GP F 15 U GT3 T 4 2 health servi… home   
## 5 GP F 16 U GT3 T 3 3 other other home   
## 6 GP M 16 U LE3 T 4 3 services other reput…  
## 7 GP M 16 U LE3 T 2 2 other other home   
## 8 GP F 17 U GT3 A 4 4 other teach… home   
## 9 GP M 15 U LE3 A 3 2 services other home   
## 10 GP M 15 U GT3 T 3 4 other other home   
## # … with 639 more rows, and 22 more variables: guardian <chr>,  
## # traveltime <dbl>, studytime <dbl>, failures <dbl>, schoolsup <chr>,  
## # famsup <chr>, paid <chr>, activities <chr>, nursery <chr>, higher <chr>,  
## # internet <chr>, romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>,  
## # Dalc <dbl>, Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>,  
## # G3 <dbl>

Data summary

|  |  |
| --- | --- |
| Name | df\_por1 |
| Number of rows | 649 |
| Number of columns | 31 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 17 |
| numeric | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

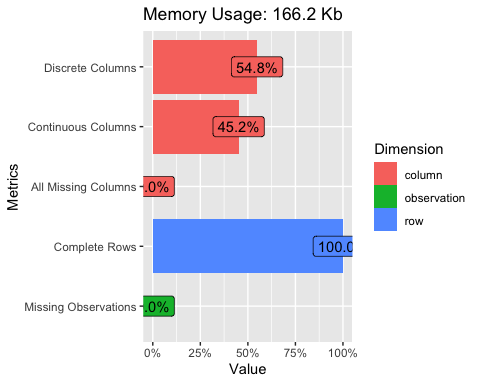
**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| school | 0 | 1 | 2 | 2 | 0 | 2 | 0 |
| sex | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| address | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| famsize | 0 | 1 | 3 | 3 | 0 | 2 | 0 |
| Pstatus | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| Mjob | 0 | 1 | 5 | 8 | 0 | 5 | 0 |
| Fjob | 0 | 1 | 5 | 8 | 0 | 5 | 0 |
| reason | 0 | 1 | 4 | 10 | 0 | 4 | 0 |
| guardian | 0 | 1 | 5 | 6 | 0 | 3 | 0 |
| schoolsup | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| famsup | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| paid | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| activities | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| nursery | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| higher | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| internet | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| romantic | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

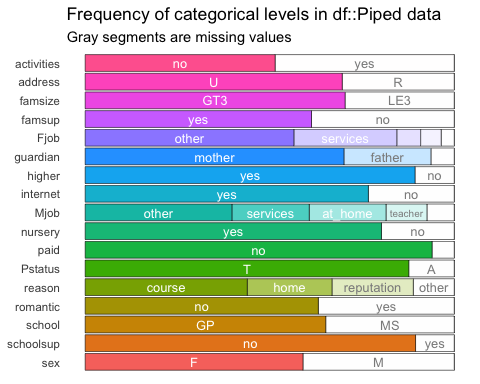
**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1 | 16.74 | 1.22 | 15 | 16 | 17 | 18 | 22 | ▇▅▅▁▁ |
| Medu | 0 | 1 | 2.51 | 1.13 | 0 | 2 | 2 | 4 | 4 | ▁▆▇▆▇ |
| Fedu | 0 | 1 | 2.31 | 1.10 | 0 | 1 | 2 | 3 | 4 | ▁▆▇▅▅ |
| traveltime | 0 | 1 | 1.57 | 0.75 | 1 | 1 | 1 | 2 | 4 | ▇▅▁▁▁ |
| studytime | 0 | 1 | 1.93 | 0.83 | 1 | 1 | 2 | 2 | 4 | ▆▇▁▂▁ |
| failures | 0 | 1 | 0.22 | 0.59 | 0 | 0 | 0 | 0 | 3 | ▇▁▁▁▁ |
| famrel | 0 | 1 | 3.93 | 0.96 | 1 | 4 | 4 | 5 | 5 | ▁▁▂▇▅ |
| freetime | 0 | 1 | 3.18 | 1.05 | 1 | 3 | 3 | 4 | 5 | ▂▃▇▆▂ |
| goout | 0 | 1 | 3.18 | 1.18 | 1 | 2 | 3 | 4 | 5 | ▂▆▇▆▅ |
| Dalc | 0 | 1 | 1.50 | 0.92 | 1 | 1 | 1 | 2 | 5 | ▇▂▁▁▁ |
| Walc | 0 | 1 | 2.28 | 1.28 | 1 | 1 | 2 | 3 | 5 | ▇▅▃▃▂ |
| health | 0 | 1 | 3.54 | 1.45 | 1 | 2 | 4 | 5 | 5 | ▃▂▃▃▇ |
| absences | 0 | 1 | 3.66 | 4.64 | 0 | 0 | 2 | 6 | 32 | ▇▂▁▁▁ |
| G3 | 0 | 1 | 11.91 | 3.23 | 0 | 10 | 12 | 14 | 19 | ▁▁▇▇▂ |

As variáveis G1 e G2 não são siguinificativas para o modelo e por isso foram retiradas do conjunto de dados. Ao analisarmos o banco de dados, observamos que ele apresenta a informação que há seisentas e quarenta e nove observações e trinta e uma variáveis, sendo com a classificação numérica (*age,Medu,Fedu,traveltime,studytime,failures,famrel,freetime,goout,Dalc,Walc,health,absences,G1,G2 e G3*) e do tipo character(*school,sex,address,famsize,Pstatus,Mjob,Fjob,reason,guardian,schoolsup,famsup,paid,activities,nursery,higher,internet e romantic*).

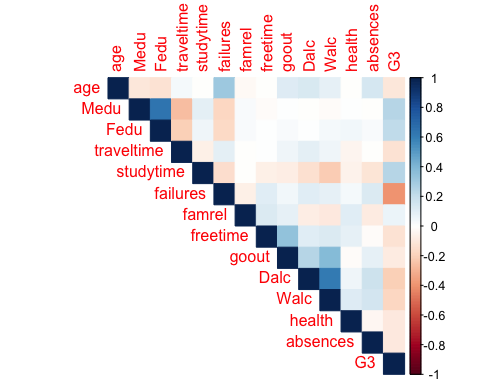


O conjunto de dados é composto por 51,8% colunas discretas, 45,2% continuas, não contém dados faltantes (NAs) e 100,00% das linhas completas.



Análise:

**5.1 Correlação**



Podemos observar que as variáveis que mais possui uma correlação moderada com nossa variável alvo (G3) são: Medu, Fedu e studytime. Com uma correlação baixa temos as variáveis Failures, Dalc e Walc.

**5.2 Análise**

Vemos claramente que temos uma tarefa de aprendizado supervisionado, uma vez que temos exemplos *rotulados* de treinamento (cada variável vem com o resultado esperado, ou seja, as notas dos alunos). Além disso, também é uma tarefa típica de regressão, já que estamos buscando prever um valor, cuja a variável alvo é do tipo quantitativa. Mais especificamente, trata-se de um problema de *regressão multivariada*, uma vez que o sistema utilizará múltiplas característica para fazer uma previsào.

**6 Regresão**

**6.1 Divisão do conjunto de dados em treinamento e teste**

Vamos separar a amostra em treino e teste.

## [1] 517 31

## [1] 132 31

Obtemos quinhentas e dezesete amostras para o nosso conjunto de treinamento e cento e trinta e duas amostras para o nosso conjunto de teste.

Iremos preparar o conjunto de treinamento para um esquema de validação cruzada com 5 folds. A **Validação Cruzada** nos permite comparar diferentes métodos de aprendizado de máquina ou parâmetros para o método escolhido e avaliar qual funcionará melhor na prática. Ou seja, o que iremos fazer é, para cada método,

1. Separar os dados em conjunto de treino e conjunto de teste.
2. Treinar um modelo no conjunto de treino.
3. Avaliar no conjunto de teste
4. Repetir os passos 1-3 e estimar o erro.

**6.2 Pré-Processamento**

Antes de criarmos um modelo de predição, é importante plotarmos as variáveis do nosso modelo antecipadamente para observarmos se há algum comportamento estranho entre elas. Por exemplo, podemos ter uma variável que assuma frequentemente um único valor (possui muito pouca variabilidade), o que não acrescenta informações relevantes ao modelo. No nosso conjunto de dados temos muitas variáveis do tipo *character* e os modelos de aprendizagem de máquina não trabalhar bem com esse tipo de dados. Resoveremos esse problema utilizando variáveis dummies. As variáveis dummies ou variáveis indicadores são formas de agregar informações qualitativas em modelos estatísticos. Ela atribui 1 se o elemento possui determinada característica, ou 0 caso ela não possua. Esse tipo de transformação é importante para modelos de regressão pois ela torna possível trabalhar com variáveis qualitativas. Iremos criar variáveis *dummy* para as variáveis qualitativas utilizando a função *step\_dummy* do pacote *tidyverse*. Nosso conjunto de dados não possui dados faltantes (NA`s), porém, aplicaremos o método *k-Nearest Neighbore (knn)* para tratar dados faltantes caso eles venham a surgir na utilização de novos dados. O método *k-Nearest Neighbore (knn)* consiste em procurar os k vizinhos mais próximos do elemento que possui o dado faltante de uma variável de interesse, calculando a média dos valores observados dessa variável dos k vizinhos e imputando esse valor ao elemento. Iremos,tambbém, normalizar as variáveis quantitativas utilizando a função *step\_normalize*. Essas transformações serão realizadas em nosso conjunto de treinamento (*df\_trn*).

**6.3 Ajustando modelos**

Nessa etapa, iremos selecionar e treinar um modelo de Aprendizado de Máquina.

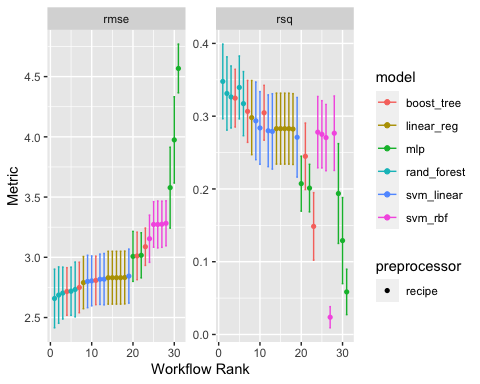
Agora, realizamos os testes dos modelos utilizando o nosso conjunto de treinamento após o seu pré-ajuste e armazenando-os na variável *wf*.

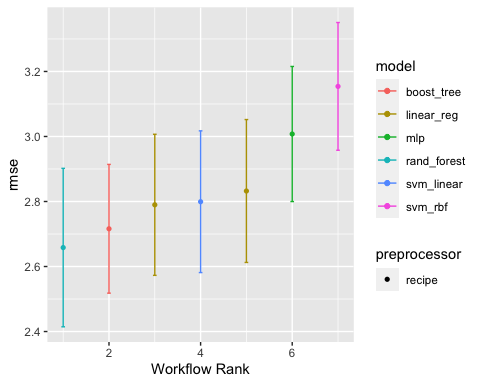
**6.4 Escolhendo o melhor modelo**

**6.4.1 Selecionando uma Medida de Desempenho**

Uma médida típica de desempenho para problemas de regressão é a *Raiz do Erro Quadrático Médio* (RMSE). Ela dá uma ideia da quantidade de erros gerados pelo sistema em nossas previsões, com um peso maior para grandes erros. O gráfico a seguir, nos mostra os calculos para duas diferentes medidas: “RMSE” e o “Rsquared”. A **Raiz do Erro Quadratico Médio** (RMSE - *Root Mean Squared Erro*), como o nome já diz, não é nada mais que a raiz quadrada do erro quadrático Médio.

Outra médida de desempenho que podemos itilizar é o **Coeficiente de Determinação**, também chamado de (R squared), é dado pela razão entre o MSE e a variância subtraído de 1.





A análise gráfica nos informa que o melhor modelo é o *random\_forest* (floresta aleatória) que apresentou uma menor medida para o *RMSE* e um maior *rsq*. Random forest (Floresta Aleatória) é um algoritmo de aprendizado de máquina que funciona de forma semelhante a uma árvore de decisão. A diferença é que a floresta aleatória usa várias árvores de decissão para fazer uma previsão e, portanto, diminui o *overfitting*. O processo de votação por maioria é realizado e a classe selecionada pela maioria das árvores é atribuída a um item.

## # A tibble: 1 × 3  
## mtry min\_n .config   
## <int> <int> <chr>   
## 1 22 16 Preprocessor1\_Model3

#ATENÇÃO conferir essa análise

O teste do *best\_results* informa e confirma a análise gráfica de que o melhor modelo é o *Preprocessor1\_Model3*, ou seja, o *random\_forest*.

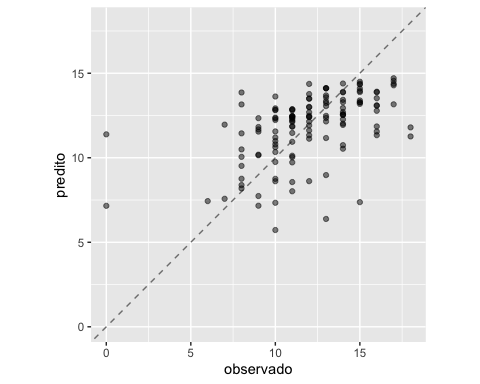
## # A tibble: 2 × 4  
## .metric .estimator .estimate .config   
## <chr> <chr> <dbl> <chr>   
## 1 rmse standard 2.63 Preprocessor1\_Model1  
## 2 rsq standard 0.263 Preprocessor1\_Model1

Assim, obtemos que o melhor modelo (*random\_forest*) apresentou um RMSE = 2.6251399 e um rsq = 0.2631965. Uma maneira de entender melhor se um determinado valor de RMSE é “bom” é normalizá-lo usando a seguinte fórmula:

Isso produz um valor entre 0 e 1, onde valores mais próximo de 0 representam modelos de melhor ajuste.

## [1] 0.1381653

Isso nos diz que o modelo é capaz de prever as notas dos alunos com precisão.



Podemos notar que há uma relação linear positiva entre os valores observados e os valores preditos.