

**UNIVERSIDADE FEDERAL DA PARAÍBA**

**CENTRO DE INFORMÁTICA**

**Relatório – Database “Student Alcohol Consumption”**

**Thiago Gonzaga Gomes – 11504760**

**Disciplina: Pesquisa Aplicada à Computação**

**Prof.ª: Thais Gaudêncio**

**João Pessoa, 27 de maio de 2016.**

**Resumo**

Esse estudo teve como objetivo abordar o vício dos estudantes quanto ao consumo de bebidas alcóolicas no ensino fundamental e médio, utilizando a técnicas de Business Intelligence (BI) e Data Mining (DM). Para a realização da pesquisa foi utilizado um conjunto de dados baseado em alunos portugueses de dois cursos (matemática e português). Os dados foram recolhidos e analisados por Paulo Cortez e Alice Silva da Universidade do Minho, Portugal. Os resultados obtidos demonstraram uma correlação entre o consumo de bebidas alcoólicas e os atributos: social, gênero e tempo de estudo para cada aluno, permitindo que uma boa precisão da previsão pudesse ser alcançada, uma vez que a dependência de álcool pode afetar diretamente o desempenho dos alunos. Desse modo, ferramentas mais precisas de previsão podem ser desenvolvidas, a fim de acompanhar e compartilhar dados a respeito do impacto do vício em bebidas alcoólicas na vida dos estudantes.

# Introdução

O consumo de bebidas alcoólicas tem um número imenso de impactos negativos na nossa vida, beber demais de uma só vez ou ao longo do tempo pode trazer consequências permanentes para a nossa saúde. As bebidas alcoólicas tem muitos efeitos de curto e longo prazo na saúde, consumidores de álcool já experimentaram alguns dos efeitos de curto prazo na saúde em algum momento de sua vida, dentre eles, a ressaca ou uma noite conturbada de sono. Pesquisas anteriores mostram o impacto do consumo de bebidas alcoólicas na adolescência, entre elas a redução das habilidades físicas e mentais, afetando assim o julgamento e a coordenação, podendo levar os adolescentes outros diversos problemas. O impacto do consumo de bebidas alcoólicas alcança valores tão altos que as funções vitais do cérebro, como o controle da respiração, pode ser bloqueado.

Alcoólatras tendem a sofrer mais acidentes e se machucar mais que pessoas que não bebem. Como aspecto mais preocupante, é possível analisar que eles tendem a se envolver em acidentes de trânsito provocados pela bebida. Quando crianças /adolescentes bebem, o sentido de julgamento e tomada de decisões é diretamente afetado, e eles tendem a sofrer grandes riscos como praticar relações sexuais sem proteção, o que pode levar a doenças sexualmente transmissíveis e gravidez indesejada. O consumo excessivo de álcool por adolescentes às vezes está ligado com outros comportamentos prejudiciais, como o uso de drogas ilícitas.

O consumo de bebidas alcoólicas pode ser uma via de sentido único, sem volta, para um sentimento de auto repulsa enorme. Menores que bebem, são mais tendenciosos a sofrerem uma série de problemas de saúde, incluindo mudanças drásticas de peso, problemas de pele, sono conturbado, dores de cabeça, além da abertura para diversos outros problemas como algumas doenças crônicas. Durante a infância e a adolescência, o cérebro ainda está em desenvolvimento, e adicionar o consumo de álcool nesse processo é problemático, já que pode afetar funções da memória, reações, sinapses nervosas, capacidade de atenção e aprendizado, que são fundamentalmente importantes nos anos escolares.

O consumo de bebidas alcoólicas pode afetar o desempenho da criança na escola a impedindo de atingir o seu potencial máximo, jovens que bebem excessivamente são mais tendenciosas a ter distúrbios mentais, até mesmo automutilação.

Todo pai quer que seu filho sempre dê o melhor de si, e o bom desempenho na escola tem uma grande influência nisso, e as estatísticas mostram que menores de idade que bebem, fazem isso com bem menos incidência. Crianças que começam a beber a partir dos 13 anos de idade, tendem a terem piores notas, saírem da escola, e no pior cenário, serem expulsas da escola. Eles têm menos autocontrole e seus cérebros não discernem os avisos de perigo, estimulando comportamentos agressivos, aumentando o envolvimento desses jovens com violência, e sérios casos de vandalismo aumentam diretamente quando comparado com o consumo de álcool, que pode levar à prisão e sujar seu nome na sociedade. Adicionar álcool nas suas bebidas não é uma boa ideia; isso pode colocá-los em situações perigosas ou invulneráveis. [1][2][3]

Nesse estudo, serão analisados dados recentes de dois colégios portugueses, onde duas fontes diferentes foram usadas: relatórios e questionários. Possibilitando assim a coleta de diversos atributos sociais e demográficos relacionados à escola. Primeiramente, foram utilizados métodos de DM (Data Mining)e árvore de decisões. Em seguida foi elaborada a análise explanatória e explicativa será feita nos melhores modelos, em busca de identificar as características mais importantes apresentada através dos dados estatísticos.

# Metodologia

Para a pesquisa foi utilizado uma data set sobre estudantes portugueses, que foi coletada por Paulo Cortez e Alice Silva da Universidade de Minho em Portugal[4]. No país, o ensino secundário consiste em três anos de escola, precedendo nove anos de educação básica e seguido do ensino superior. A maioria dos estudantes ingressam no sistema público e gratuito de ensino. Esse estudo considerou dados coletados durante os anos escolares de 2005-2006 de duas escolas públicas, da região portuguesa de Alentejo (centro-sul de Portugal).

A database utilizada foi construída a partir de duas fontes: relatórios escolares, baseados em documentos que incluem alguns atributos e questionários, usados para complementar as informações anteriores. Os questionários foram projetados com perguntas relacionadas a diversas variáveis demográficas (e.g. educação dos pais, renda familiar), social/emocional (e.g. consumo de álcool, relacionamentos) e escolares (e.g. número de reprovações) que eram esperadas para ter efeito direto na performance do estudante.

Os atributos são mostrados abaixo na Tabela 1, onde as últimas quatro linhas denotam as variáveis retiradas dos relatórios escolares. Para cada atributo, existe uma descrição e o valor assumido por ele, seja binário, nominal ou numérico. O impacto de cada atributo foi extraído, analisado e estudado.

**Tabela 1**

***Pré-processamento: variáveis relacionadas com os alunos***

|  |  |
| --- | --- |
| **Atributo** | **Descrição** |
| sex | sexo do estudante (binário: **M** homem ou **F** mulher) |
| age | idade do estudante (numérico: de **15** a **22**) |
| School | escola do estudante (binário: **GP** *Gabriel Pereira* ou **MS** *Mousinho da Silveira*) |
| address | endereço do estudante (binário: **U** urbano ou **R** rural) |
| famsize | tamanho da família (binário: **LE3 ≤ 3**, **GT3 > 3**) |
| Pstatus | coabitação dos pais (binário: **T** vivendo juntos ou **A** separados) |
| Medu/Fedu | educação da mãe (numérico: de **0** a **4**) |
| Mjob/Fjob | emprego da mãe (nominal: **perto de casa**, **reputação da escola**, ou **outro**) |
| reason | razão por escolher a escola (nominal: perto de ‘**casa**’, ‘**reputação**’ da escola, ‘**curso**’ ou ‘**outro**’ |
| guardian | responsável pelo estudante (nominal: **mother** mãe, **father** pai, **other** outro. |
| traveltime | tempo de casa até a escola (numérico: **1** <15min, **2** 15 a 30 min, **3** 30 min a 1 hora ou **4** >1 hora) |
| studytime | tempo de estudo semanal (numérico: **1<** 2 horas, **2** 2 a 5 horas, **3** 5 a 10 horas ou **4 >**10 horas) |
| failures | número de reprovações (numérico: **n** se **1 ≤ n < 3**, caso contrário 4) |
| schoolsup | suporte extra educacional da escola (binário: **sim** ou **não**) |
| famsup | suporte educacional da família (binário: **sim** ou **não**) |
| activities | atividades extra curriculares (binário: **sim** ou **não**) |
| paidclass | aulas extras pagas (binário: **sim** ou **não**) |
| internet | acesso à internet em casa (binário: **sim** ou **não**) |
| nursery | foi à enfermaria na escola (binário: **sim** ou **não**) |
| higher | almeja o ensino superior (binário: **sim** ou **não**) |
| romantic | tem um relacionamento sério (binário: **sim** ou **não**) |
| famrel | qualidade do relacionamento c/ família (numérico: de **1** muito ruim a **5** muito bom) |
| freetime | tempo livre após a escola (numérico: de **1** pouco a **5** muito) |
| goout | saindo com amigos (numérico: de **1** pouco a **5** muito) |
| health | estado atual de saúde (numérico: de **1** muito ruim a **5** muito bom) |
| absenses | número de faltas na escola (numérico: de **1** a **93**) |
| dalc | consumo em dia de expediente (numérico: de **1** pouco a **5** muito) |
| walc | consumo em finais de semana (numérico: de **1** pouco a **5** muito) |
| G1 | nota no primeiro período (numérico: de **1** a **20**) |
| G2 | nota no segundo período (numérico: de **1** a **20**) |
| G3 | nota final (numérico: de **1** a **20**) |
| alc | consumo de álcool durante a semana + dia de expediente (numérico: de **1** pouco a **5** muito) |

# Pré-processamento

Databases reais são altamente suscetíveis a ruídos, dados inconsistentes e falta de informações devido ao seu tamanho tipicamente enorme (geralmente com gigabytes de tamanho) e sua propensão de origem de várias, heterogêneas fontes. Dados de baixa qualidade geralmente tendem a resultados de baixa qualidade. Desse modo, várias técnicas de pré-processamento podem ser usadas[5] para realizar um tratamento de dados a fim de remover os ruídos e corrigir as inconsistências nos dados e fundir duas databases diferentes usando integração de dados. Existem infinitos motivos para imprecisão de dados (i.e., valores de atributos incorretos), no entanto, as ferramentas para coleta dos dados podem ser falhas.

Vários algoritmos de Data Mining foram utilizados, cada um com seus propósitos e suas capabilidades, foram almejados para as tarefas de classificação. Além das técnicas de DM, a árvore de decisões também foi implementada, cuja característica é ser uma estrutura de ramificação que representa um conjunto de regras, diferindo valores em forma hierárquica [6]. Essa representação pode ser traduzida em um conjunto de regras IF-THEN (SE-ENTÃO), que podem ser facilmente ser entendidas por pessoas.

Vale salientar que podem existir erros de entrada dos dados, tanto humanos como da própria máquina. Os usuários podem propositalmente submeter valores incorretos dos dados dos campos obrigatórios quando querem ocultar informações pessoais (e.g., escolhendo o valor padrão “1 de janeiro” exibido para data de aniversário), esse tipo de erro é conhecido como dado camuflado.

As duas datasets, uma contendo as informações dos estudantes do curso de matemática, e a outra contendo as informações dos estudantes do curso de português, são extraídas, lidas e unidas, pelo script:

|  |
| --- |
| >d1=read.table("student-mat.csv",sep=";",header=TRUE) >d2=read.table("student-por.csv",sep=";",header=TRUE)  > total <- rbind(d1,d2) |

O objetivo foi achar a média de consumo alcóolico por estudantes do ensino secundário. Nessa database, existem dois atributos diferentes para o álcool.

O primeiro é o consumo de bebida alcóolica em um dia normal de trabalho/expediente (Dalc) e o outro é o consumo de bebida alcoólica no final de semana (Walc). Só um índice pode ser previsto, com isso foi feito um atributo que representa o consumo total de álcool por um estudante específico em uma semana inteira. A união dos dois atributos é feita pela seguinte equação:

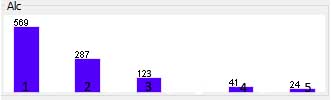
Onde ***Walc*** é multiplicado por 2, devido à quantidade de dias presentes no final de semana (sábado e domingo) e ***Dalc*** multiplicado por 5, devido à quantidade de dias presentes nos dias de trabalho/expediente (segunda à sexta).  
O novo atributo, ***Alc***, muda entre um e cinco (pouco a muito). Com intuito de checar se um estudante é um consumidor ou não. ***Alc*** se torna um valor binário. Se ***Alc < 3***, o binário assume o valor de 0, significando que o estudante em caso não é um consumidor, caso contrário, ***Alc > 3***, o binário assume o valor de 1 e o estudante é um consumidor.

Adiciona-se o valor de ***Alc***, no nosso dataset com o script:

|  |
| --- |
| >format(round(Alc<-total[,((Walc\*2)+(Dalc\*5))/7]), nsmall=2)  >total[,"Alc"]<-c(Alc) |

Observamos no gráfico 1, a distribuição das 1044 amostras

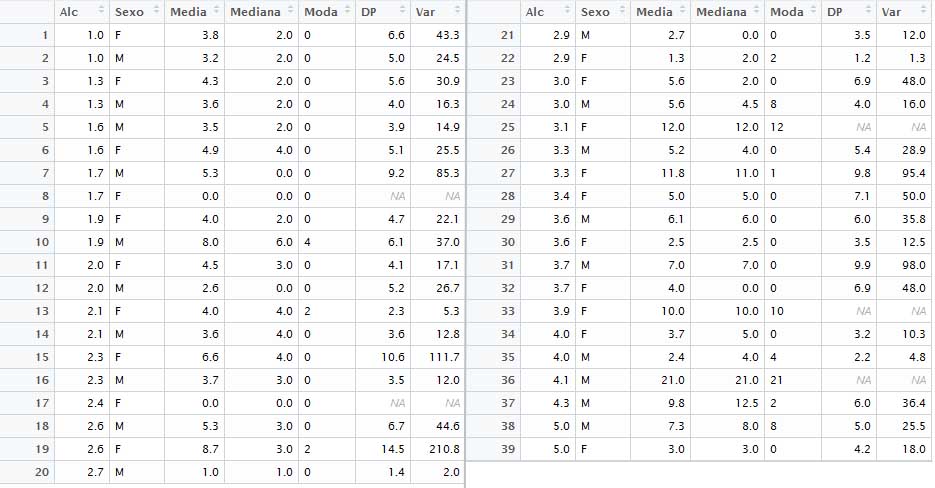
*Gráfico 1*



Por meio do script, utilizado no RStudio:

|  |
| --- |
| > ind1<-total[,.(Media=mean(absences), Mediana=quantile(absences,0.5), Moda=names(which.max(table(absences))), DP=sd(absences), Var=var(absences)), by=.('Alc'=Alc, 'Sexo'=sex)]  > setorder(ind1, Alc) |

Podemos obter um índice contendo a análise do número de faltas escolares por consumo semanal de bebidas alcoólicas (Alc) por meio da Média, Mediana, Moda, Desvio Padrão e Variância.



*\*Ruídos nos itens 15, 19, 27 e 31 não identificados\**

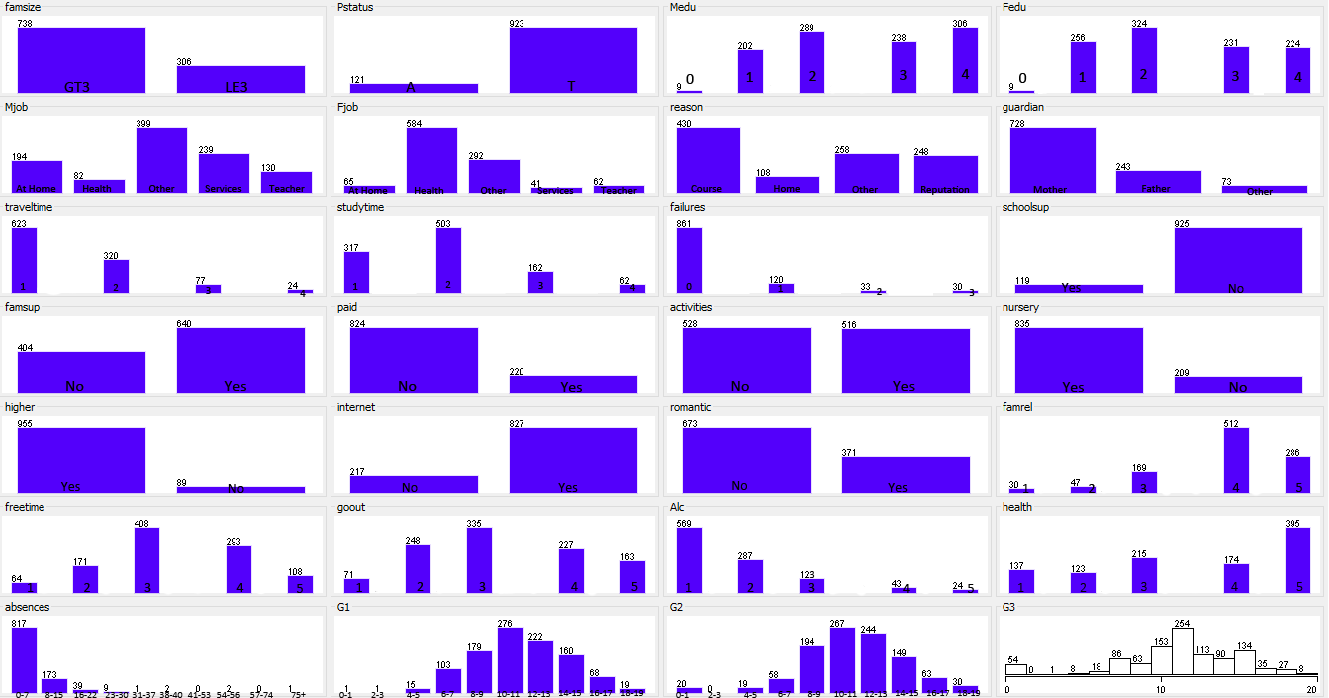
Por meio do script:

|  |
| --- |
| > with(total,cor(absences,Alc)) |

Podemos obter a correlação entre o número de faltas e o consumo de bebidas alcóolicos, que resulta numa taxa de 0.15%. Com a mesma função, podemos realizar outras várias correlações entre atributos numéricos.

Plotando um gráfico de *barplot* para todos os atributos do dataset, no *Gráfico 2* conseguimos quantificar as 1044 amostras para cada contexto relacionado.

*Gráfico 2*

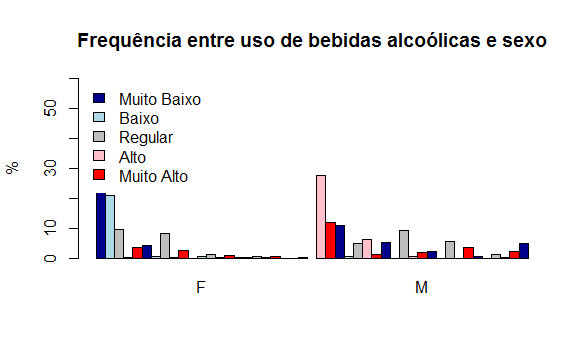


Analisando a taxa de faltas do aluno, tem-se que se um estudante falta aulas frequentemente, ele tem mais tempo livre, com isso, consume mais bebidas alcóolicas. Assim, o atributo torna-se binário, se o aluno falta com frequência (mais de 10 dias), o valor é 0, caso contrário, o valor é 1.

  
*Gráfico 3*



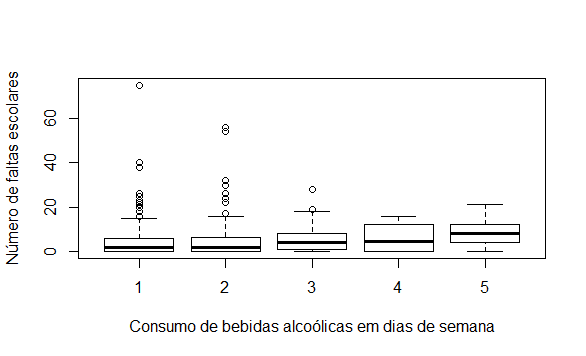
Podemos também, averiguar graficamente a incidência da correlação entre o consumo de bebidas alcoólicas (Alc) e o gênero do estudante.  
*Gráfico 4*



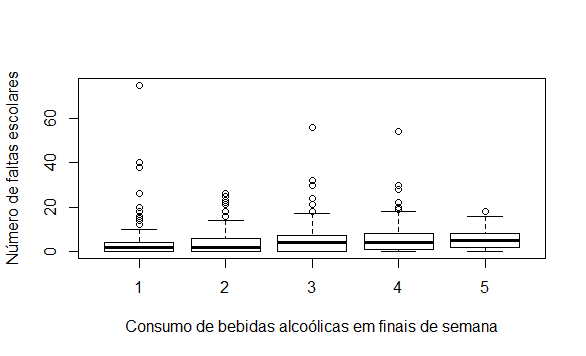
Constata-se a maior incidência e maior frequência de consumo pelo sexo masculino.

Fazendo a comparação entre as faltas escolares e os consumos de bebida alcóolica (Dalc, Walc e Alc) é possível constatar os ruídos por meio de boxplot.

*Gráfico 5. Comparação entre o número de faltas e o consumo de bebidas alcóolicas em dias de semana.*



*Gráfico 6. Comparação entre o número de faltas e o consumo de bebidas alcóolicas em finais de semana.*



# Resultados e Discussão

Após o término do pré-processamento, os atributos de maior impacto na determinação do nível de alcoolismo do estudante foram escolhidos. Ao total, foram 25 identificadores, que auxiliaram a representar/traçar o perfil do estudante consumidor. A porcentagem do valor do impacto é mostrada na *Tabela 2* e no *Gráfico 4*.

**Tabela 2**

***Atributos mais relevantes/impactantes para o estudante consumidor***

|  |  |
| --- | --- |
| **Atributo** | **Porcentagem** |
| Homem | 25.35% |
| Social | 21.13% |
| Mais tempo livre | 9.39% |
| Menos tempo de estudo | 8.45% |
| Menor qualidade na educação da mãe | 7.98% |
| Boa saúde | 7.04% |
| Não almeja ensino superior | 4.23% |
| Sem/pouco suporte da família | 3.76% |
| Família pequena | 3.76% |
| Longo tempo de viagem até a escola | 1.88% |
| Poucas atividades | 1.88% |
| Sem/pouco suporte da escola | 1.88% |
| Trabalho do pai | 1.88% |
| Conectividade com a internet | 1.41% |

*Gráfico 7*

A porcentagem da relevância é calculada com o uso da equação:

Onde, **A** é um conjunto de atributos, ***l*** representa o número do nível (alto ou baixo), ***T*** é o conjunto de árvores de decisão (estrutura de ramificação que representa um conjunto de regras, diferindo valores em forma hierárquica), **C** conta o número de vezes que o atributo está presente em uma árvore de atributos em um nível específico e **6 – i** representa o peso do atributo.

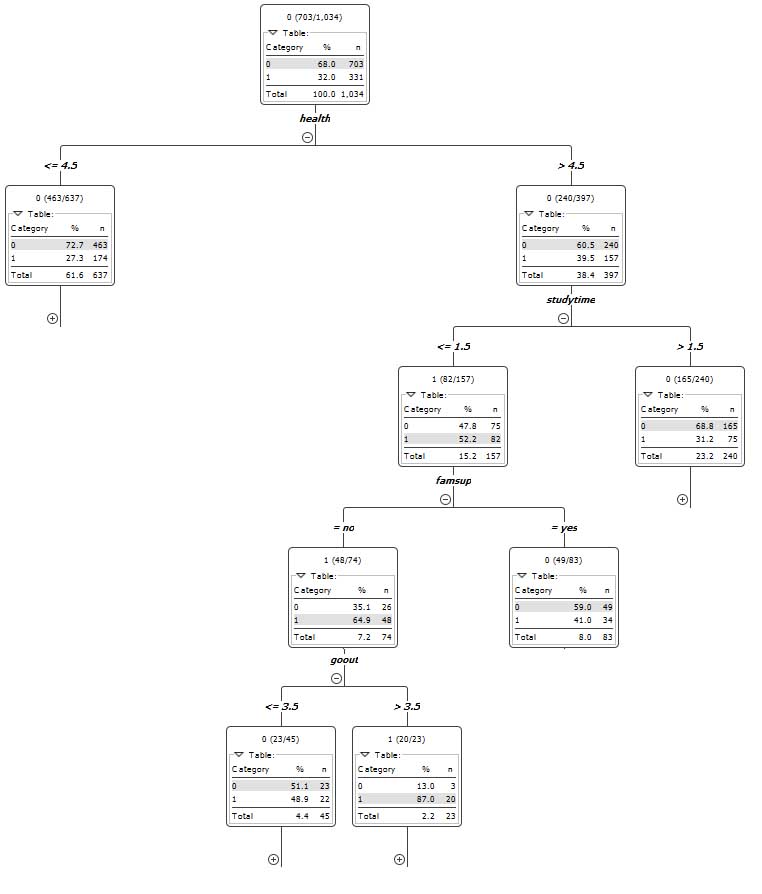
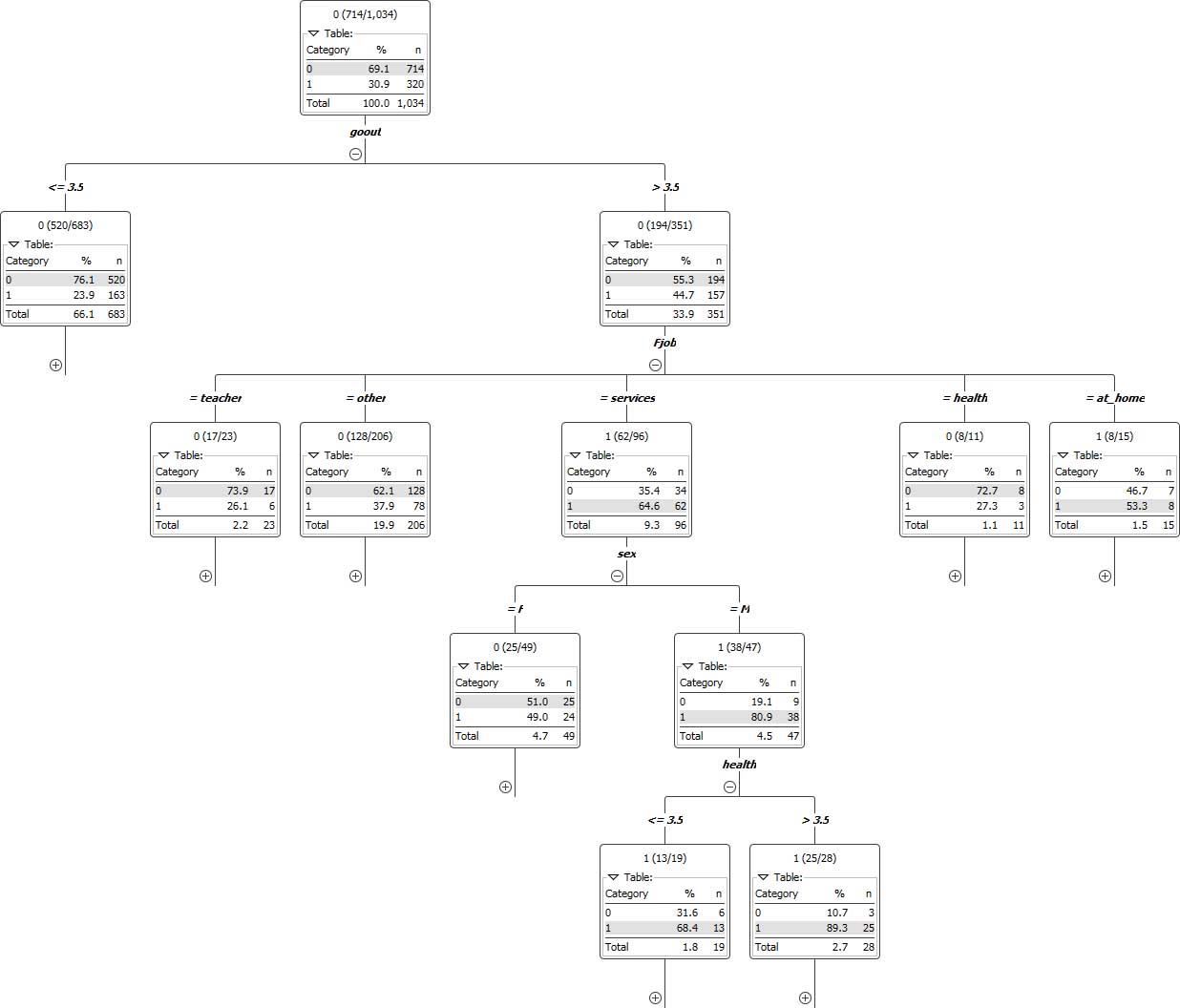
A partir da análise do gráfico e tabela, como anteriormente exposto, percebeu-se que o gênero (masculino) é a característica de maior peso para um consumidor, indicando que homens são os mais envolvidos com bebidas alcóolicas. Esse resultado vai de acordo com os dados da Organização Mundial da Saúde (OMS)[7]. Segundo a OMS, homens bebem consideravelmente mais que mulheres. A pessoa que sai mais frequentemente com os amigos, consome mais álcool. Isso se dá por conta de que beber, em nossa sociedade, tornou-se uma forma de celebração e também, pela indução de que pessoas bêbadas são mais “disponíveis”. Algumas outras características importantes que levam ao consumo são: mais tempo livre e menos tempo de estudo, tendo em vista que, geralmente, se encontra na bebida alcóolica, uma forma de relaxar e/ou escapar dos problemas.

O alvo principal nesse estudo foram os adolescentes, os mesmos irão decidir seu próprio futuro: trabalho ou universidade. E de acordo com os resultados e estatísticas, quem não frequentar a universidade, irá consumir mais bebidas alcóolicas. Essa característica é correlacionada de alguma forma com o tempo gasto com estudos.

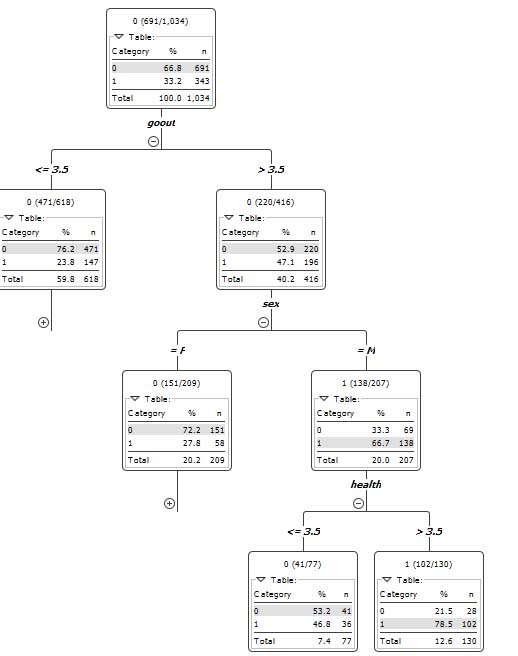
Um fator que não pode ser esquecido é a família. Apoio para estudo, um grande número de pessoas na família e um bom emprego dos pais são fatores importantes no crescimento de uma criança. Tampouco como a educação da mãe é algo mais que fundamental na vida da criança. O estudo mostra que a falta ou escassez de algumas dessas características levam a criança a ter mais chances de aderir à vida alcóolatra.

Por meio das árvores de decisão, consegue-se fazer um diagrama de previsão das causas/razões/circunstâncias pelas quais o jovem encaminha-se para o consumo de bebidas alcoólicas. Vide *Figuras 1,2,3 e 4.*

*Árvore de Decisões 1*

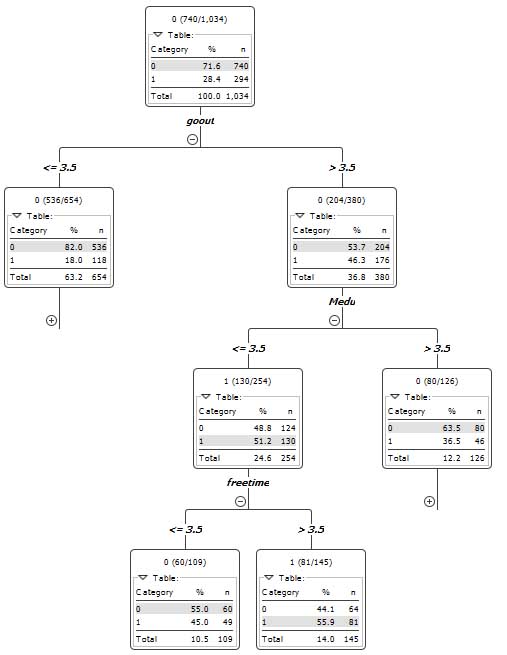


*Árvore de Decisões 2*



*Árvore de Decisões 3.*

*Árvore de Decisões 4.*



Além disso, mais atividades e maior tempo de viagem entre a casa e a escola são uma fonte de estresse, uma vez que isso implica sedentarismo e pouco tempo de descanso para essa pessoa. Outro grande problema chegando no último século, internet. Isso pode influenciar qualquer pessoa de uma forma negativa, tendo em vista que pode ser utilizada da forma errada. Algumas combinações entre características podem ser perigosas e aumentam a chance da criança se tornar viciada em bebidas alcóolicas. Nas figuras das árvores de decisões 1,2,3 e 4, podemos observar alguns dos fatores e sua correlação com a taxa dos estudantes viciados em álcool.

# Conclusão

A educação é um elemento crucial na nossa sociedade. Técnicas de Business Intelligence (BI) e Data Mining (DM), as quais permitem uma extração de grande nível de conhecimento de dados sem tratamento, oferecem possibilidades muito interessantes para o domínio educacional. Em particular, muitos estudos têm usados métodos de BI/DM para reduzir o nível de vício em álcool para adolescentes e melhorar o estilo de vida para as crianças. Nesse trabalho, foram endereçadas predições de vício alcóolico com adolescentes por meio de registros escolares, e dados demográficos, familiares, sociais e outros relacionados ao estudante. Alguns métodos e objetivos de DM foram testados. Os resultados obtidos revelam que é possível manter a criança longe do álcool. Isso confirma a conclusão encontrada que o comportamento da criança é altamente afetado pelos amigos ou pelo seu círculo social. Todavia, uma análise mais profunda com esse conhecimento mostrado pelos melhores métodos, mostra que, em alguns casos, existem outros fatores relevantes, tais como: variáveis sociais, demográficas e relacionadas à escola.

Este trabalho foi baseado em um estudo prévio, de um recolhimento e junção de uma database recolhida por professores da Universidade de Minho, Portugal. Porém, existe um imensurável conteúdo online que aumenta o ambiente de pesquisa quando se trata de alcoolismo com menores de idade.

# Referências

[1] Drinkaware.co.uk. Why underage drinking is a risky business.

[2] Drugfreeworld.org. The truth about alcohol.

[3] National Institute on Alcohol Abuse and Alcohlism (NIH). Alcohol’s effects on the body.

[4] 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 (*fubutec2008*) pp:5-12 porto portugal]. 2008.

[5] Jiawei Han Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, third edition.

[6] Breiman L. Friedman J. Ohlsen R. and Stone C. classification and regression trees. 1984.

[7] Com (*Srl*). Villars sous Yens. Global status report on alcohol and health. Technical report, World Health Organization, 2014.

Chambers, J.M. (2008) Software for data analysis: Programming with R.

De'Ath, G. & Fabricius, K.E. (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, 81, 3178-3192.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2011) R package version 1.17-6.

Quick-R. Disponível em: <<http://www.statmethods.net/>> Acesso em: 24/05/16.

Cookbook for R. Disponível em: < [http://www.cookbook-r.com/](%20http://www.cookbook-r.com/)> Acesso em: 24/05/16.

Stack Overflow. Disponível em: < http://stackoverflow.com/> Acesso em: 26/05/16.