Pré-Processamento de Dados Físico-Químicos de Vinhos

Matheus Gomes Cordeiro[1] Abel Pinheiro de Figueiredo[2] Michelly Karen Diógenes Pereira[3] Rodrigo Silva Lima[4]

Universidade Federal do Ceará (UFC) - Departamento de Engenharia de Teleinformática (DETI)
60430-160, Fortaleza - Brasil

Resumo. Ao estudar uma quantidade significativa de amostras de vinhos portugueses (mais especificamente, o Vinho Verde branco) Cortez et al. [1] modelou as relações entre características físico-químicas e a qualidade percebida, o que pode auxiliar no trabalho de avaliadores de vinhos e na percepção das preferências de seus compradores. Ao se realizar o pré-processamento dos dados do dataset gerado, procurou-se analisar e representar as relações entre as diferentes características dos vinhos e suas respectivas classificações de uma forma mais clara e objetiva, gerando histogramas, valores de média e outros dados estatisticamente relevantes, para facilitar seu uso em estudos posteriores.

1 Introdução

O vinho faz parte de um mercado bem lucrativo em muitos países da Europa, incluso nesse grupo, Portugal se destaca por seu famoso Vinho do Porto e também por seu Vinho Verde. Nesse sentido, a certificação de vinho se mostra como uma tarefa de extrema complexidade, pois além de testes físico-químicos são feitos testes sensoriais, e relacionar propriedades materiais de um composto com algo que tange a subjetividade é muito difícil. Atrelada a certificação está a avaliação de qualidade que auxilia na produção de vinho, identificando fatores mais influentes, o que permite a direcionar o cultivo e o processo industrial para vinhos específicos de preço já definido.

Motivado por trabalhos passados onde o número de amostras era bem menor e a literatura continha apenas um trabalho relacionado a qualidade de vinhos, Cortez et al. [1], utilizando um conjunto de amostras bem robusto, procurou modelar a preferência, a qual embasa a qualidade, por meio de dados analíticos oriundos do processo de certifição de vinhos, mais especificadamente, de Vinhos Verdes. Isso possui uma gama de utilidades, duas muito pertinentes são: a ajuda ao processo decisão dos enólogos e o auxílio na modelagem das preferências de compradores de vinho, a segunda se destaca no campo do *Business* pois entendendo o gosto do cliente é possível desenvolver um tipo de vinho que busca conquistar setores de mercado mais lucrativos.

Esse trabalho se dispôs a realizar o pré-processamento de dados de um dos datasets de Cortez et al. [1] (dataset de Vinho Verde branco), visando um entendimento mais aprofundado de como o conjunto de dados está estruturado e de como as características constituíntes do vinho (variáveis) se relacionam com as respectivas notas de qualidade (classes) e também entre sí. Um estudo mais aprofundado como esse é importante, pois esclarece detalhes do dataset que podem ser relevantes para um possível preditor no futuro, almejando um viés de aplicação dentro do contexto de Cortez et al. [1] no âmbito da análise de preferência.

2 Metodologia

Utilizando o dataset de Vinhos Verdes brancos abordado anteriormente, o qual foi baixado do UCI repository [2], procurou-se entender sua estrutura básica a princípio. Esse conjunto de dados possui 11 variáveis independentes: Acidez Volátil, Acidez Fixa, Ácido Cítrico, Residual de Açucar, Cloreto, Dióxido de Enxofre Livre, Dióxido de Enxofre Total, Densidade, pH, Sulfatos e Álcool, e possui como variável dependente a qualidade do vinho que é representada por números em

um intervalo inteiro fechado de 0 a 10 (notas). Existem 4898 amostras de Vinho Verde branco distribuídas da seguinte forma: 0 de qualidade 0, 0 de qualidade 1, 0 de qualidade 2, 20 de qualidade 3, 163 de qualidade 4, 1457 de qualidade 5, 2198 de qualidade 6, 880 de qualidade 7, 175 de qualidade 8, 5 de qualidade 9 e 0 de qualidade 10, tal que essas notas são obtidas pelo seguinte método: após a análise das propriedades físico-químicas do vinho, a amostra é submetida ao crivo de três avaliadores e então a mediana das três notas é atribuída como nota da amostra [1]. Após esse estudo inicial, fazendo o uso da linguagem **R** [3] dentro do ambiente de desenvolvimento RStudio [4] e utilizando os respectivos pacotes: ggplot2 [5], factorextra [3] e corrplot [3] foram executados os seguintes procedimentos [6]:

2.0.1 Análise Monovariada e Não-Condicional dos Dados

Os histogramas da distribuição de cada variável do dataset foram plotados, dos quais os mais importantes foram os de Dióxido de Enxofre Livre (Figura 3), Residual de Açucar (Figura 2), e Acidez Fixa (Figura 1), porque apresentam maiores desvios padrão, além do de Cloreto (Figura 4), pois o mesmo apresenta a maior assimetria (o Dióxido de Enxofre Total possui um desvio padrão maior que o do Dióxido de Enxofre Livre, entretando ele depende diretamente do Dióxido de Enxofre Livre (Figura 3). Logo em seguida foi feita uma tabela (Tabela 1) com as médias, os desvios padrão e os valores de skewness de cada variável. Cada uma dessas métricas possui sua importância em um âmbito analítico: a média de uma variável, tecnicamente, fornece o valor com a maior probabilidade de ser encontrado, no dataset, daquela variável (esperança matemática), o desvio padrão, por sua vez, quantifica o grau de dispersão dos valores de uma variável aleatória em torno da média, ou seja, um desvio padrão alto indica que há amostras muito afastadas da média, já a skewness informa a obliquidade da distribuição, a mesma varia positivamente e negativamente, se ela for muito maior que zero configura uma assimetria positiva o que significa que existem muito mais valores acima da média, o contrário ocorre quando a skewness é negativa, e quando ela é próxima de zero ou igual a zero a distribuição é caracterizada como simétrica, o que implica muitos valores próximos da média.

Variáveis	Média	Desvio Padrão	Skewness
Acidez Volátil	6.854	0.843	0.647
Acidez Fixa	0.278	0.100	1.576
Ácido Cítrico	0.334	0.121	1.281
Açucar Residual	6.391	5.072	1.076
Cloreto	0.045	0.021	5.020
Dióxido de Enxofre Livre	35.308	17.007	1.406
Dióxido de Enxofre Total	138.361	42.498	0.390
Densidade	0.994	0.002	0.977
pH	3.188	0.151	0.457
Sulfato	0.489	0.114	0.976
Alcool	10.514	1.230	0.487

Tabela 1: Analise Não-condicional

2.0.2 Análise Monovariada e Classe-Condicional dos Dados

Os histogramas de frequência de todas as variáveis para cada classe além da tabela de médias, desvios padrão e skewness por classe foram construídos. Os mais relevantes foram, pela mesma

justificativa da subseção anterior, os de Residual de Açucar e de Dióxido de Enxofre Livre, e na espectativa de avaliar o comportamento de cada variável, segundo o crecimento da qualidade, foram escolhidas classes específicas a serem destacadas: Residual de Açucar: Classe 4 (Figura 5), Classe 6 (Figura 6) e Dióxido de Enxofre Livre: Classe 4 (Figura 7), Classe 6 (Figura 8).

2.0.3 Análise Bivariada e Não-Condicional dos Dados

Os gráficos de dispersão entre cada par de variável além da matriz de correlação (Figura 9) entre as variáveis foram plotados. Os gráficos mais importantes foram os de Densidade vs Residual de Açucar (Figura 10) e Álcool vs Desnsidade (Figura 11), pelo fato das variáveis presentes neles possuírem as maiores correlções diretas e inversas, segundo a Matriz de Correlação (Figura 9) .

2.0.4 Análise das Componentes Principais

Uma Análise das Componentes Principais (PCA) do dataset foi feita sendo plotados um gráfico de barras mostrando a variância por componente (somente as duas primeiras componentes) (Fig. 4.0) e um gráfico de dispersão das duas primeiras componentes onde as classes são representadas pelas cores dos pontos (Fig. 4.1). O PCA é, resumidamente, uma uma transformação ortogonal que converte o plano em que se encontram os dados para um plano cujos eixos são os preditores de maior variância. Ele é relevante pois o mesmo consegue reduzir a dimensionalidade do dataset e assim permite uma análise mais refinada do comportamento das variáveis independentes.

3 Resultados

Desse modo, diante do apresentado é possível chegar as seguintes conclusões:

3.0.1 Sobre as Propriedades Físico-Químicas Estudadas

Há uma possibilidade de existirem outliers abaixo da média na Acidez Fixa da Classe 4 [6] e na Acidez Volátil de todas as classes (Tabela 1), no caso da segunda devido ao seu sekewness elevado (skewness = 1.576). Isso implica que as amostras de vinho escolhidas apresentam, em sua maioria, um gosto bastante avinagrado (azedo), já que, abaixo de 0.4 g(ácido acético)/dm3 (Tabela 1), o gosto azedo se acentua, além disso a Classe 4 possui uma amostra que excede 0.9 g(ácido acético)/dm3, isso é bastante intrigante pois esse valor é classificado como impróprio para o consumo, assim como alguns valores de Acidez Total da Classe 4 e da Classe 5, já que chegam perto de 3.5 g(ácido tartárico)/dm3 [6].

Dando continuidade, o Residual de Açucar possui uma relação muito íntima com a taxa de Álcool, devido ao fato de que durante a fermentação do vinho as leveduras convertem o açucar em álcool dentro de seu processo de obtenção de energia, caso ocorra algum problema nesse procedimento a taxa de açucar que sobra (Açucar Residual) pode ser alta ou baixa, implicando, inversamente, na taxa de Álcool. Nesse contexto, a densidade se mostra como um fator intermediário que traduz essa relação, já que quanto maior for a densidade de um vinho maior será sua taxa de Açucar Residual e menor será sua taxa de Álcool, algo que pode ser explicado pela massa molecular do Açucar e do Álcool. Assim os gráficos de dispersão de Densidade vs Residual de Açucar (Figura 10), Álcool vs Densidade (Figura 11) e a matriz de correlção (Figura 9) são explicados, inclusive esses gráficos de disperção apresentam linearidade.

3.0.2 Sobre o Conjunto de Dados Estudado

O dataset múito denso, ou seja, as classes não possuem fronteira muito bem definida como mostra a Figura 12, principalmente, as classes 5, 6 e 7, as quais apresentam maior quantidade de amostras, além disso o PCA mostrou que as maiores componentes principais não conseguem ter sozinhas uma boa variação dos preditores (Var(PC1) = 29,3%, Var(PC2) = 14,3%) [6]. Isso pode ser justificado pelo fato de ser muito complicado relacionar os gostos dos avaliadores, com a qualidade e com os constituíntes químicos de um vinho, logo a classes são bem próximas

3.0.3 Observações Finais

As variáveis restantes também possuem peculiaridades que merecem ser exploradas, as mesmas contém correlações bem pertinentes e visíveis na matriz de correlação (Figura 9), entretanto é algo que sai do escopo desse artigo. Por fim, todos os códigos e gráficos não apresentados nesse trabalho estão na referência: [6].

Referências

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Modeling wine preferences by data mining from physicochemical properties, *In Decision Support Systems*, Elsevier, 47(4):547-553, 2009.

[2] A. Asuncion, D. Newman, UCI Machine Learning Repository, University of California, Irvine, 2007, http://www.ics.uci.edu/ mlearn/MLRepository.html.

[3] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-00-3, 2008, http://www.R-project.org.

[4] RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL http://www.rstudio.com/.

[5] Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.

[6] Cordeiro M.G, Pré-Processamento de Dados Físico-Químicos de Vinhos, Homework 1 (2018), repositório GitHub: https://github.com/matheus123deimos/Machine-Learning.

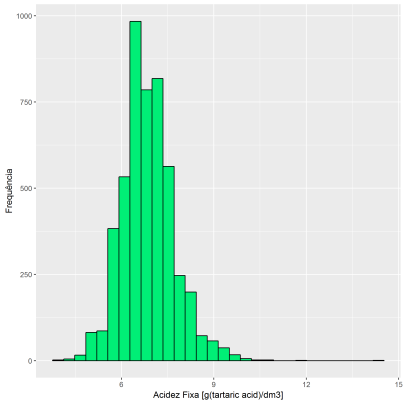

Figura 1: Acidez Fixa

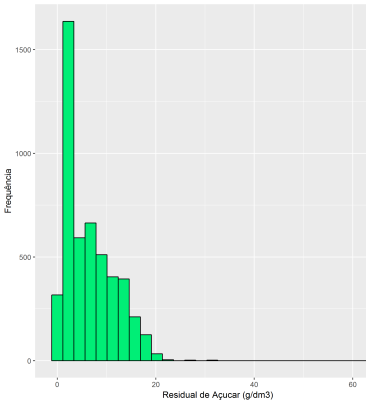

Figura 2: Resid de Açucar

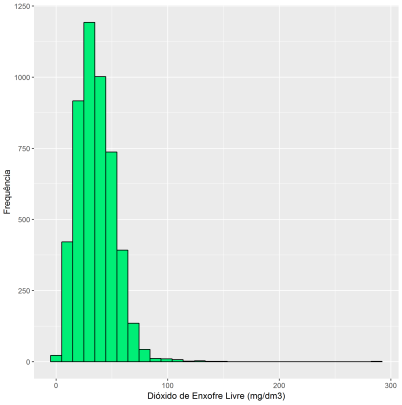

Figura 3: SO2 Livre

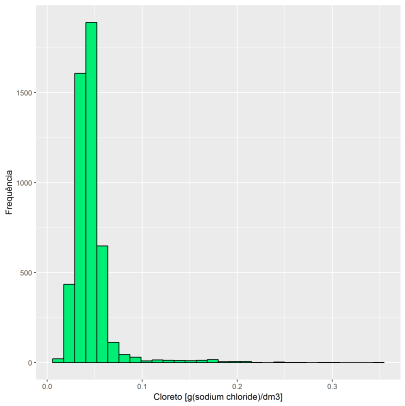

Figura 4: Cloreto

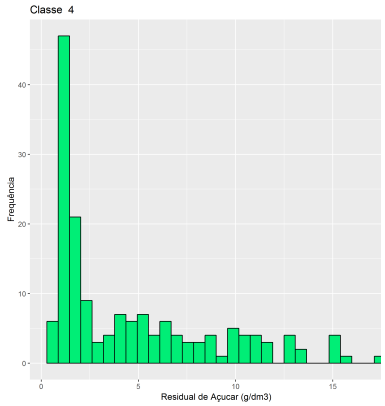

Figura 5: Clss4 Resid Açucar

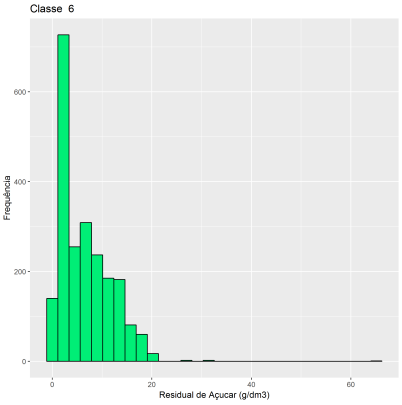

Figura 6: Clss6 Resid Açucar

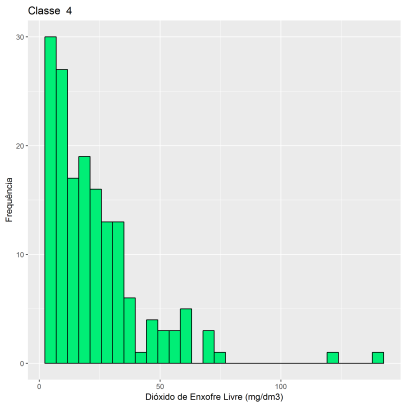

Figura 7: Clss4 SO2 Livre

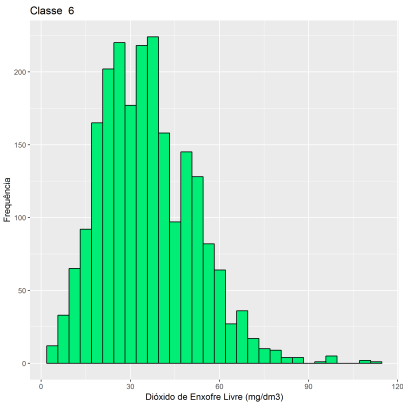

Figura 8: Clss6 SO2 Livre

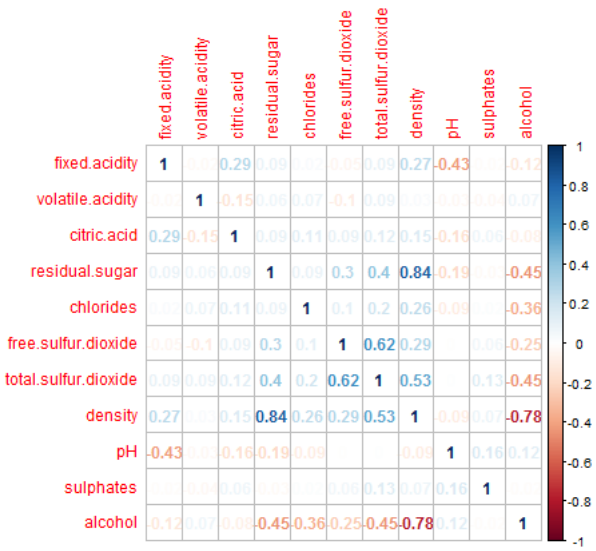

Figura 9: Matriz de Correlação

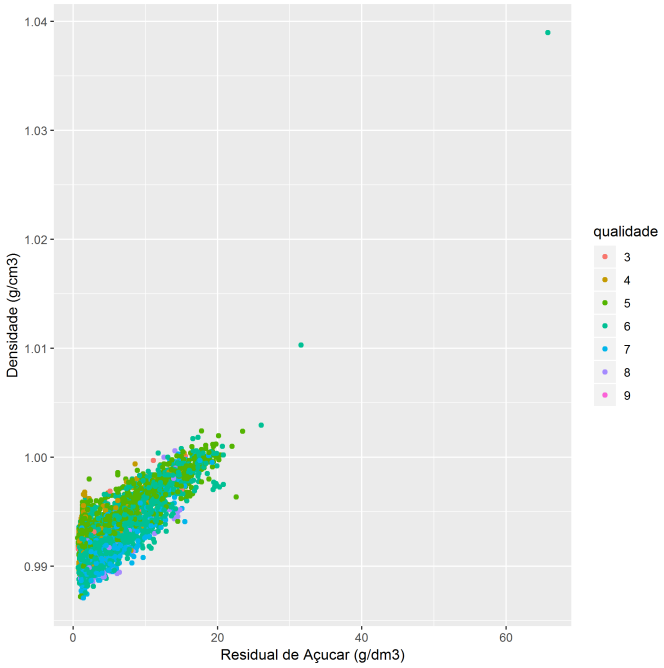

Figura 10: Densidade vs ResdAçu

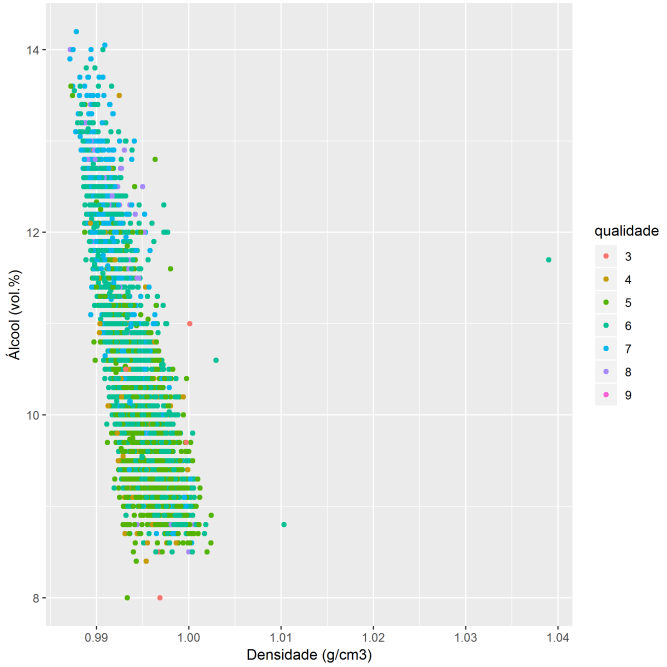

Figura 11: Álcool vs Densidade

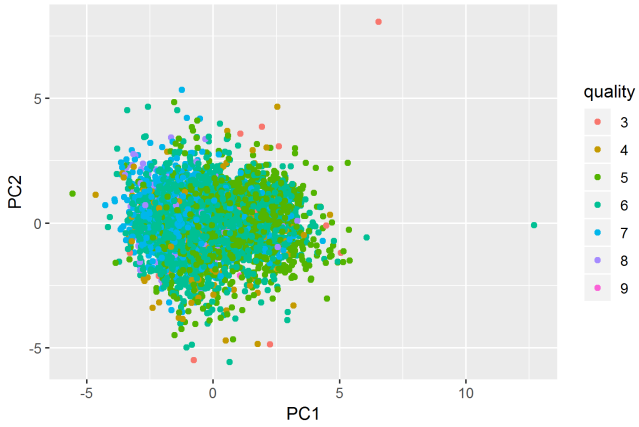

Figura 12: PC1 vs PC2