

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Ericsson Gracioli

Análise e Exploração dos Dados (EDA) para investigar como características químicas influenciam a nota de qualidade de vinhos

Belo Horizonte

2019

Ericsson Gracioli

Análise e Exploração dos Dados (EDA) para investigar como características químicas influenciam a nota de qualidade de vinhos

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2019

SUMÁRIO

1. Introdução	5
1.1. Contextualização	5
1.2. O problema proposto	5
2. Coleta de Dados	7
3. Processamento/Tratamento de Dados	9
4. Análise e Exploração dos Dados	10
5. Criação de Modelos de Machine Learning	16
6. Apresentação dos Resultados	18
7. Links	21

1. Introdução

1.1. Contextualização

A avaliação que diferentes tipos de vinhos recebem está relacionada com dois grupos de características, um deles voltado para as questões sensoriais, como aspecto visual, paladar e olfato e outro, voltado para as características químicas, extraídas por meio de análises laboratoriais. Mas haveria uma forma de relacionar, ou mesmo inferir, quais são as principais características químicas que podem influenciar a nota de qualidade que um vinho irá receber?

Iremos utilizar técnicas de classificação, extração e análise de dados para explorar a questão levantada anteriormente, bem como tentaremos elaborar modelos de inteligência artificial para predizer qual é a nota de qualidade que um vinho poderá receber a partir de suas características químicas.

1.2. O problema proposto

Para ajudar a explorar de forma mais ampla a questão central da análise de dados proposta foi elaborada a seguinte tabela utilizando a técnica dos [5-Ws](#).

Why? (Porque?) <i>Por que esse problema é importante?</i>	A análise das características químicas do vinho pode apoiar a tomada de decisão desde o produtor, até o consumidor final, passando pela cadeia de revendedores inclusive, fazendo que eles optem por um determinado rótulo, em detrimento de outro.
Who? (Quem?) <i>De quem são os dados analisados? De um governo? Um ministério ou secretaria? Dados de clientes?</i>	Os dois conjuntos de dados que utilizaremos são disponibilizados no repositório público de dados da UCI. E tem sua autoria creditada ao trabalho de <i>Cortez et al.</i>
What? (O que?) <i>Quais os objetivos com</i>	Iremos analisar como várias características químicas podem se relacionar com a nota de qualidade que é atribuída a um

essa análise? O que iremos analisar?	vinho, dentro de dois conjuntos de amostras: vinhos tintos e vinhos brancos.
Where? (Onde?) Trata dos aspectos geográficos e logísticos de sua análise	Os conjuntos de dados referem-se a amostras de vinhos verdes, brancos e tintos, da região do Minho em Portugal.
When? (Quando?) Qual o período está sendo analisado? A última semana? Os últimos 6 meses? O ano passado?	Os conjuntos de dados não possuem um período de referência estabelecido. A informação que está na página da UCI é que eles foram doados para a comunidade em: 07/10/2009. Na referência [1] é mencionado que o período de coleta foi de maio/2004 a fevereiro/2007.

2. Coleta de Dados

Os conjuntos de dados utilizados na análise foram obtido no repositório público da UCI e são acessíveis pelo link: <https://archive.ics.uci.edu/ml/datasets/Wine%2BQuality>

Na descrição dos conjuntos de dados é relatado que a coleta pode ser creditada a *Cortez et al.* e que os conjuntos foram utilizados na produção do artigo: <https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub>

Abaixo temos uma tabela que descreve cada uma das variáveis dos conjuntos de dados, lembrando que os dois conjuntos de dados compartilham as mesmas variáveis.

Nome da coluna/campo	Descrição	Tipo
fixed acidity (tartaric acid - g / dm ³)	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)	Number (double)
volatile acidity (acetic acid - g / dm ³)	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste	Number (double)
citric acid (g / dm ³)	found in small quantities, citric acid can add 'freshness' and flavor to wines	Number (double)
residual sugar (g / dm ³)	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet	Number (double)
chlorides (sodium chloride - g / dm ³)	the amount of salt in the wine	Number (double)
free sulfur dioxide (mg / dm ³)	the free form of SO ₂ exists in equilibrium between molecular SO ₂ (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine	Number (double)
total sulfur dioxide (mg / dm ³)	amount of free and bound forms of SO ₂ ; in low concentrations, SO ₂ is mostly undetectable in wine, but at free SO ₂ concentrations over 50 ppm, SO ₂ becomes evident in the nose and taste of wine	Number (double)
density (g / cm ³)	the density of water is close to that of water depending on the percent alcohol and sugar content	Number (double)

pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale	Number (double)
sulphates (potassium sulphate - g / dm ³)	a wine additive which can contribute to sulfur dioxide gas (SO ₂) levels, which acts as an antimicrobial and antioxidant	Number (double)
alcohol (% by volume)	the percent alcohol content of the wine	Number (double)
quality (score between 0 and 10)	Nota de qualidade atribuída a cada uma das amostras dos conjuntos de dados.	Number (integer)

3. Processamento/Tratamento de Dados

Como os conjuntos de dados em questão foram produzidos e utilizados em outras análises acadêmicas e explorações, já estavam bem consistentes, sem a presença de valores duplicados, ou ausentes.

Abaixo podemos verificar as estatísticas de cada um dos conjuntos de dados:

```

fixed.acidity  volatile.acidity  citric.acid  residual.sugar  chlorides
Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900
Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747
3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000
Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100

free.sulfur.dioxide  total.sulfur.dioxide  density  pH
Min.   : 1.00   Min.   : 6.00   Min.   :0.9901   Min.   :2.740
1st Qu.: 7.00   1st Qu.: 22.00   1st Qu.:0.9956   1st Qu.:3.210
Median :14.00   Median : 38.00   Median :0.9968   Median :3.310
Mean   :15.87   Mean   : 46.47   Mean   :0.9967   Mean   :3.311
3rd Qu.:21.00   3rd Qu.: 62.00   3rd Qu.:0.9978   3rd Qu.:3.400
Max.   :72.00   Max.   :289.00   Max.   :1.0037   Max.   :4.010

sulphates  alcohol  quality
Min.   :0.3300   Min.   : 8.40   Min.   :3.000
1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median :0.6200   Median :10.20   Median :6.000
Mean   :0.6581   Mean   :10.42   Mean   :5.636
3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.   :2.0000   Max.   :14.90   Max.   :8.000

```

Resumo (função summary) do dataset de vinhos tintos

fixed.acidity	volatile.acidity	citric.acid	residual.sugar
Min. : 3.800	Min. : 0.0800	Min. : 0.0000	Min. : 0.600
1st Qu.: 6.300	1st Qu.: 0.2100	1st Qu.: 0.2700	1st Qu.: 1.700
Median : 6.800	Median : 0.2600	Median : 0.3200	Median : 5.200
Mean : 6.855	Mean : 0.2782	Mean : 0.3342	Mean : 6.391
3rd Qu.: 7.300	3rd Qu.: 0.3200	3rd Qu.: 0.3900	3rd Qu.: 9.900
Max. : 14.200	Max. : 1.1000	Max. : 1.6600	Max. : 65.800

chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
Min. : 0.00900	Min. : 2.00	Min. : 9.0	Min. : 0.9871
1st Qu.: 0.03600	1st Qu.: 23.00	1st Qu.: 108.0	1st Qu.: 0.9917
Median : 0.04300	Median : 34.00	Median : 134.0	Median : 0.9937
Mean : 0.04577	Mean : 35.31	Mean : 138.4	Mean : 0.9940
3rd Qu.: 0.05000	3rd Qu.: 46.00	3rd Qu.: 167.0	3rd Qu.: 0.9961
Max. : 0.34600	Max. : 289.00	Max. : 440.0	Max. : 1.0390

pH	sulphates	alcohol	quality
Min. : 2.720	Min. : 0.2200	Min. : 8.00	Min. : 3.000
1st Qu.: 3.090	1st Qu.: 0.4100	1st Qu.: 9.50	1st Qu.: 5.000
Median : 3.180	Median : 0.4700	Median : 10.40	Median : 6.000
Mean : 3.188	Mean : 0.4898	Mean : 10.51	Mean : 5.878
3rd Qu.: 3.280	3rd Qu.: 0.5500	3rd Qu.: 11.40	3rd Qu.: 6.000
Max. : 3.820	Max. : 1.0800	Max. : 14.20	Max. : 9.000

Resumo (função summary) do dataset de vinhos brancos

Para facilitar o processo de análise e diminuir a complexidade, optei em criar uma nova variável chamada 'avaliação', derivada de 'qualidade', assim restringindo as possibilidades de nota que um vinho pode receber para: 'baixa', 'média' e 'alta'.

```

49
50 #Criaremos uma variavel categorica para refletir os valores atribuidos a
51 #variavel quality e facilitar as proximas analises.
52 vinho_tinto$avaliacao <- ifelse(vinho_tinto$quality <= 4, 'baixa', ifelse(
53   vinho_tinto$quality <= 6, 'média', 'alta'))
54
55 vinho_branco$avaliacao <- ifelse(vinho_branco$quality <= 4, 'baixa', ifelse(
56   vinho_branco$quality <= 6, 'média', 'alta'))
57
58 #Ordenando pela variavel avaliacao
59 vinho_tinto$avaliacao <- ordered(vinho_tinto$avaliacao,
60   levels = c('baixa', 'média', 'alta'))
61
62 vinho_branco$avaliacao <- ordered(vinho_branco$avaliacao,
63   levels = c('baixa', 'média', 'alta'))
64

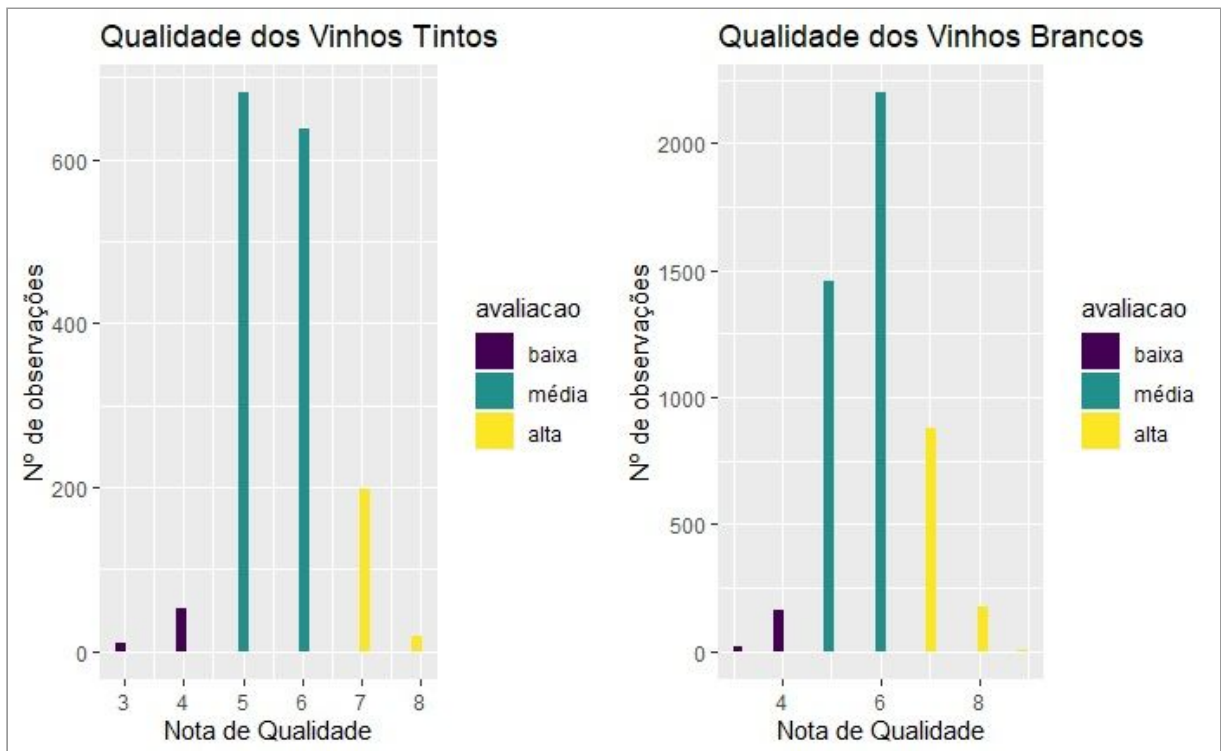
```

Resumo (função summary) do dataset de vinhos brancos

Já para a etapa de criação dos modelos de Machine Learning, que será abordada na seção 5, optei em não utilizar a variável 'avaliação', conduzindo a análise com os conjuntos de dados no estado original.

4. Análise e Exploração dos Dados

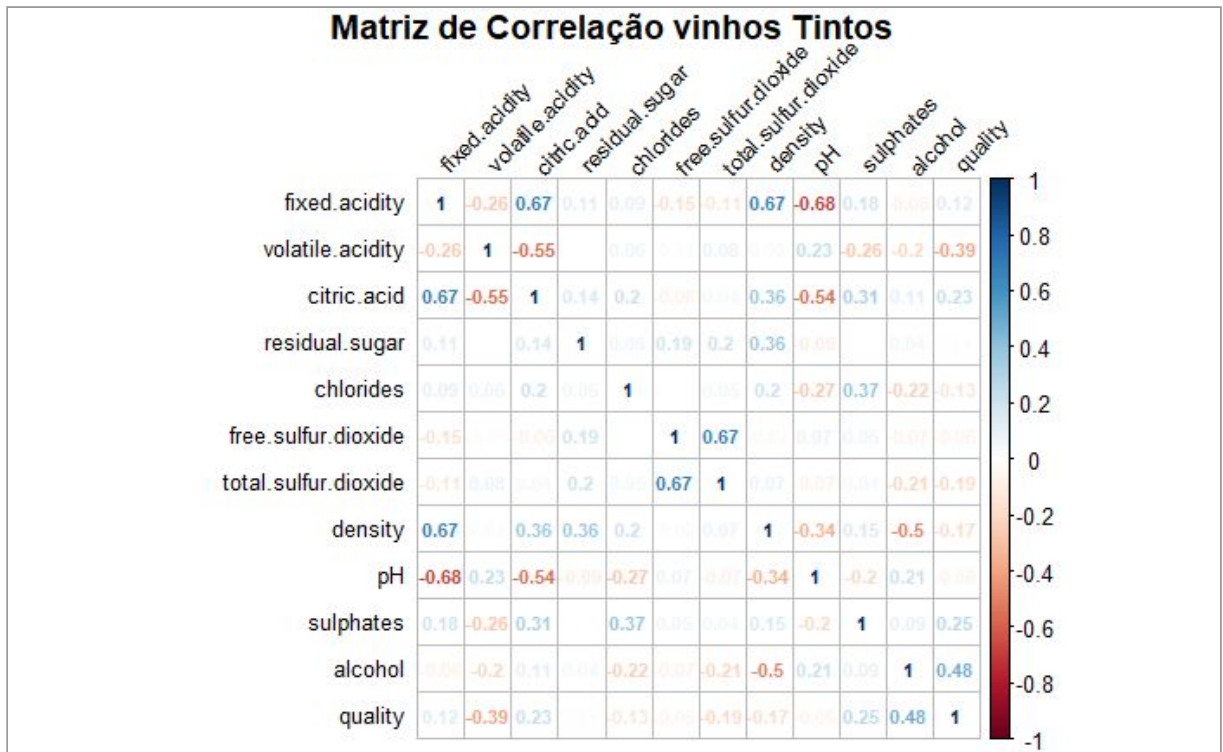
Inicialmente podemos verificar pelo histograma abaixo qual é a distribuição das 6497 amostras nos conjuntos de dados. Notamos que apesar do conjunto de dados dos vinhos brancos ser maior, ambos os conjuntos possuem mais amostras de 'avalição' igual a 'média'.



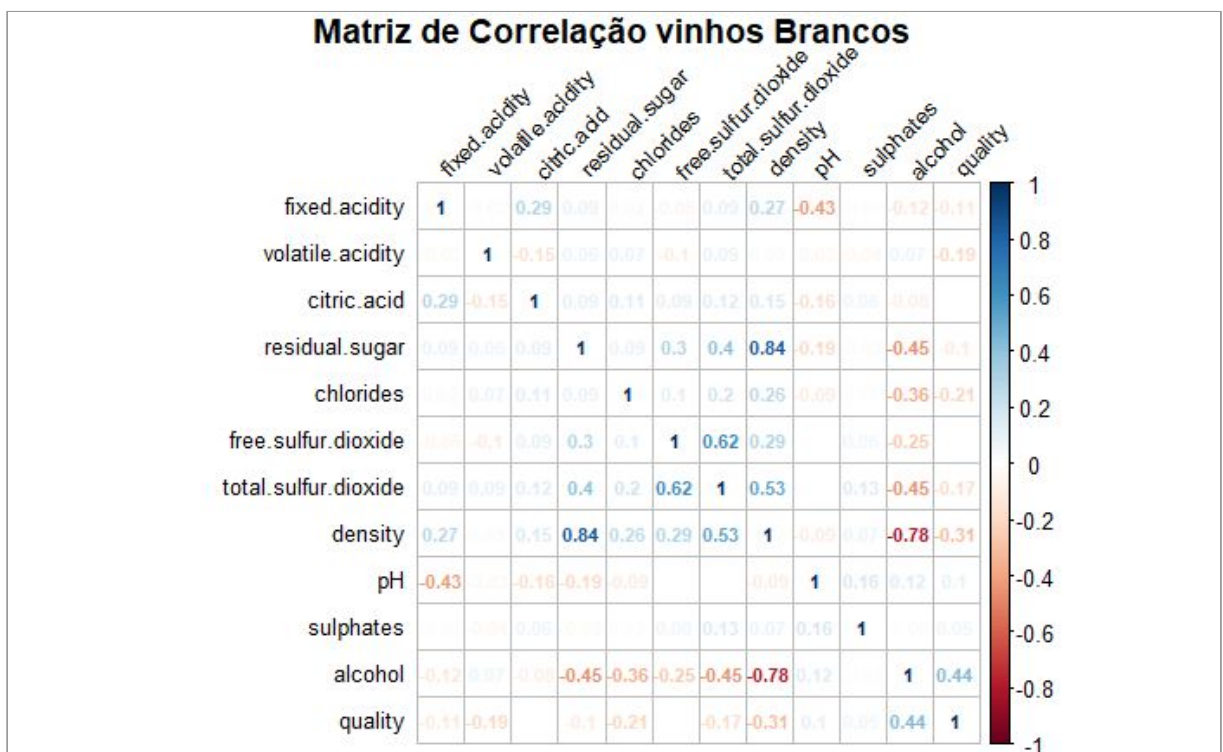
Histograma das amostras de vinho de acordo com a nota de qualidade recebida

Com a criação das matrizes de correlação dos conjuntos de dados podemos identificar as variáveis mais correlacionadas com a variável 'qualidade', no caso:

- para o conjunto de dados de vinhos tintos: são as variáveis 'alcohol' (0,48) e 'volatile.acidity' (-0,39).
- para o conjunto de dados de vinhos brancos: são as variáveis 'alcohol' (0,44) e 'density' (-0,31).

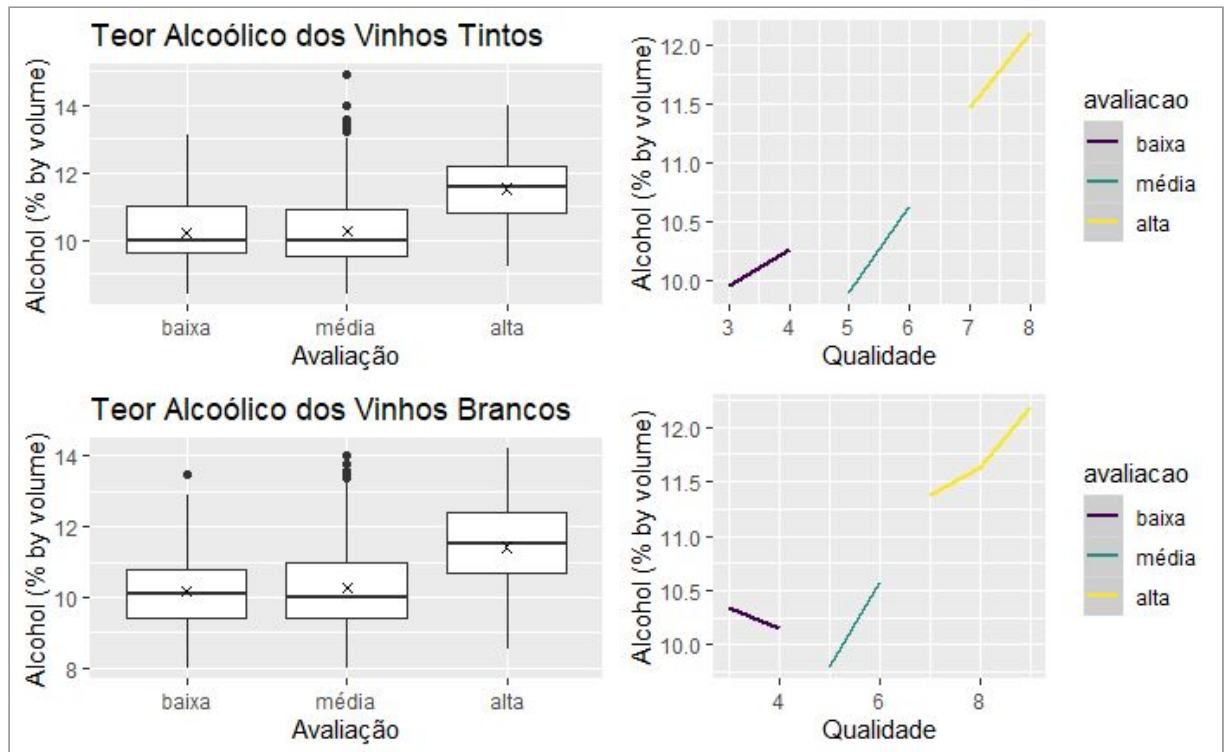


Matriz de correlação entre as variáveis do conjunto de dados de vinhos tintos



Matriz de correlação entre as variáveis do conjunto de dados de vinhos brancos

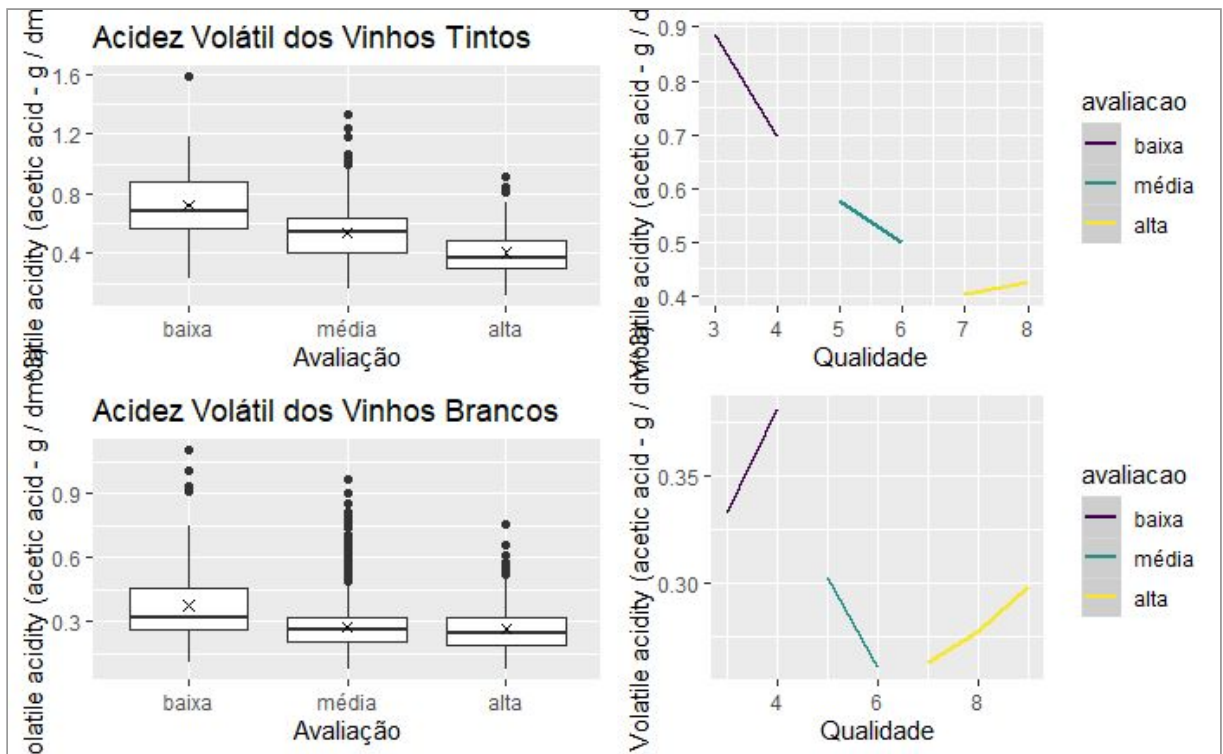
Como existe uma correlação significativa entre as variáveis alcohol e quality para ambos os tipos de vinho, podemos analisar a distribuição das observações entre as categorias de Avaliação. E verificamos que existe uma diferenciação do teor alcoólico dos vinhos com uma avaliação mais alta.



Comparação do teor alcoólico entre os conjunto de dados de vinhos tintos e brancos

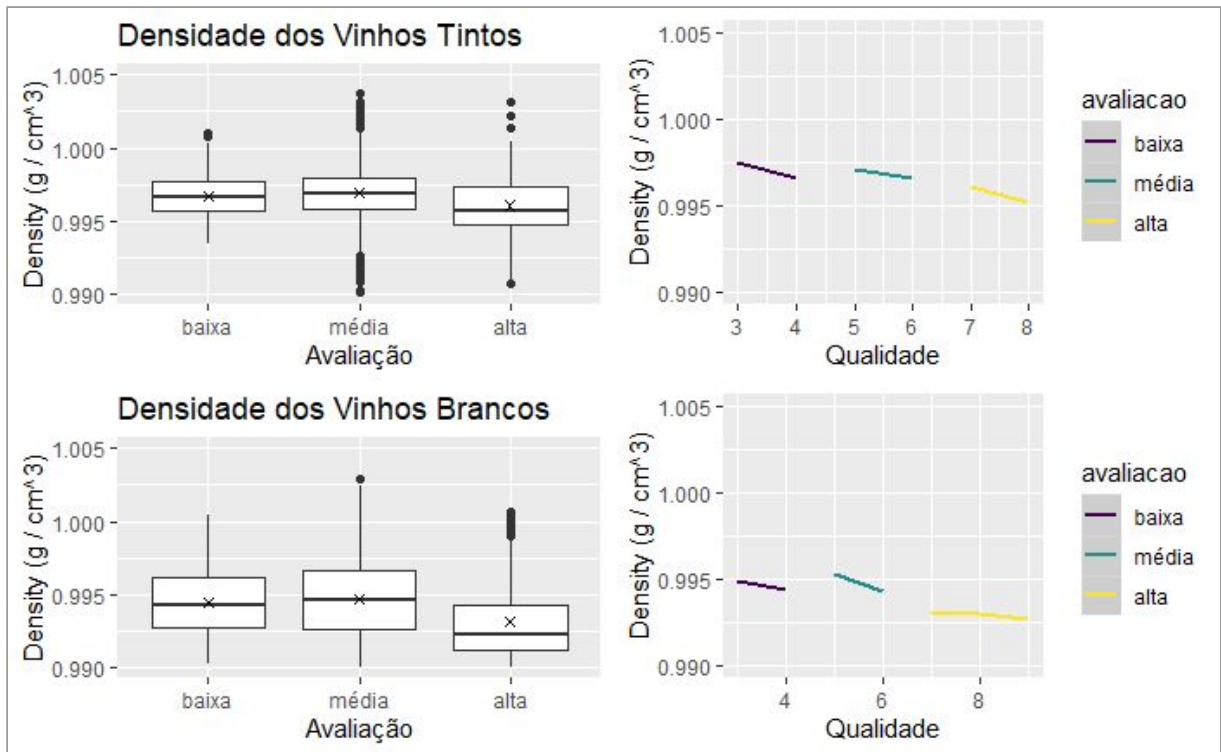
Também existe uma correlação significativa entre as variáveis volatile.acidity e quality para os vinhos tintos e que não se repete para os vinhos brancos.

Para os vinhos tintos, observamos que a volatile.acidity cai de acordo com que as notas de qualidade aumentam. E para os vinhos brancos há baixa influência da acidez volátil na nota de qualidade.



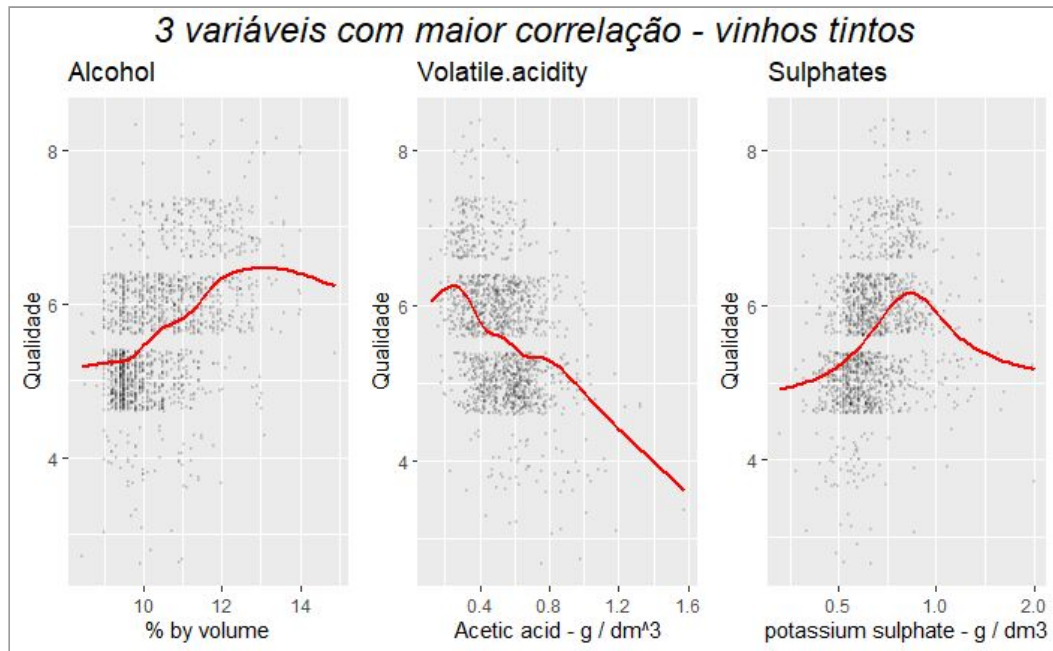
Comparação da acidez volátil entre os conjunto de dados de vinhos tintos e brancos

Já para o conjunto de dados de vinho branco, há uma correlação identificável entre as variáveis density e quality. Assim, podemos analisar a distribuição das observações entre as categorias de Avaliação. Infelizmente como a correlação entre as variáveis não é tão elevada, não foi possível identificar com ênfase no gráfico a segmentação entre classes de avaliação.



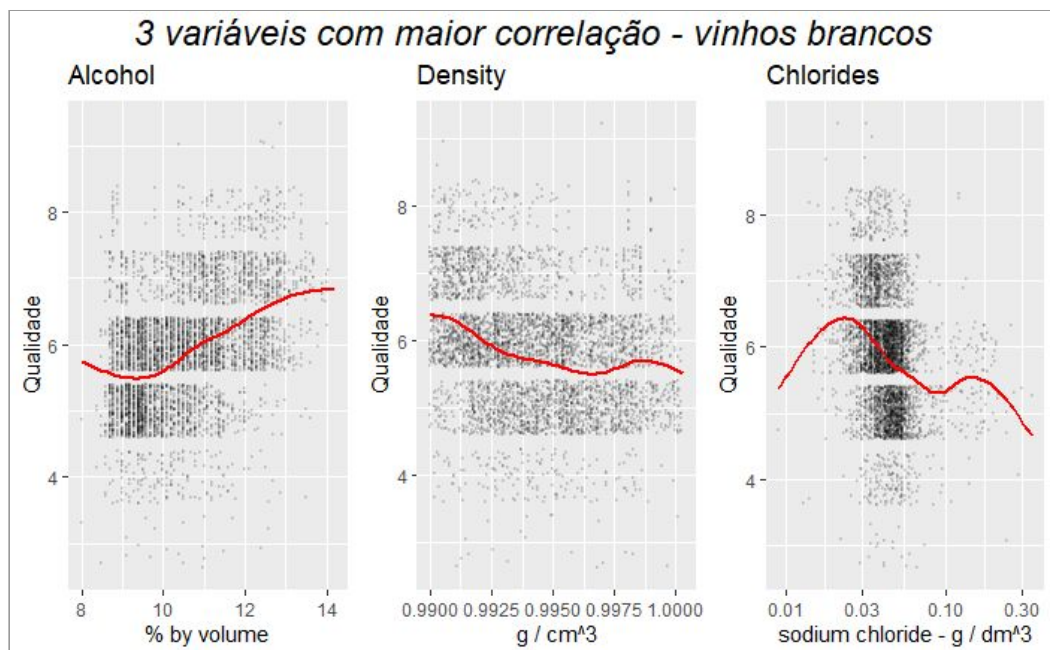
Comparação da densidade entre os conjunto de dados de vinhos tintos e brancos

Após analisarmos cada uma das correlações de forma individualizada, podemos construir uma visão consolidada entre elas. E verificamos o comportamento de cada uma delas com o aumento das notas de qualidade, onde alcohol tem uma variação positiva, volatile.acidity negativa e sulphates segue uma distribuição normal, no caso dos vinhos tintos.



Comportamento das três variáveis de maior correlação com a nota de qualidade no conjunto de dados de vinhos tintos

No caso dos vinhos brancos, observamos que o aumento das notas de qualidade, acompanha uma variação positiva para o teor alcoólico, negativa para densidade e cloretos.



Comportamento das três variáveis de maior correlação com a nota de qualidade no conjunto de dados de vinhos brancos

5. Criação de Modelos de Machine Learning

Optei por conduzir a minha análise de dados verificando qual técnica atingiria a maior acurácia em ambos os conjuntos de dados, fazendo uso da ferramenta Knime, onde decidi pelas seguintes parametrizações em todos os modelos:

- Criar cinco validações nos conjuntos de dados, ou seja, foram construídas cinco partições estratificadas a partir da variável 'qualidade' para diversificar os dados com o intuito de explicar a nota de qualidade a partir das características químicas dos vinhos.
- Separar as cinco partições aleatórias mais uma vez para obter os dados de treinamento e os dados de teste, já que o modelo não deve ser treinado com os dados que serão utilizados para a sua avaliação.

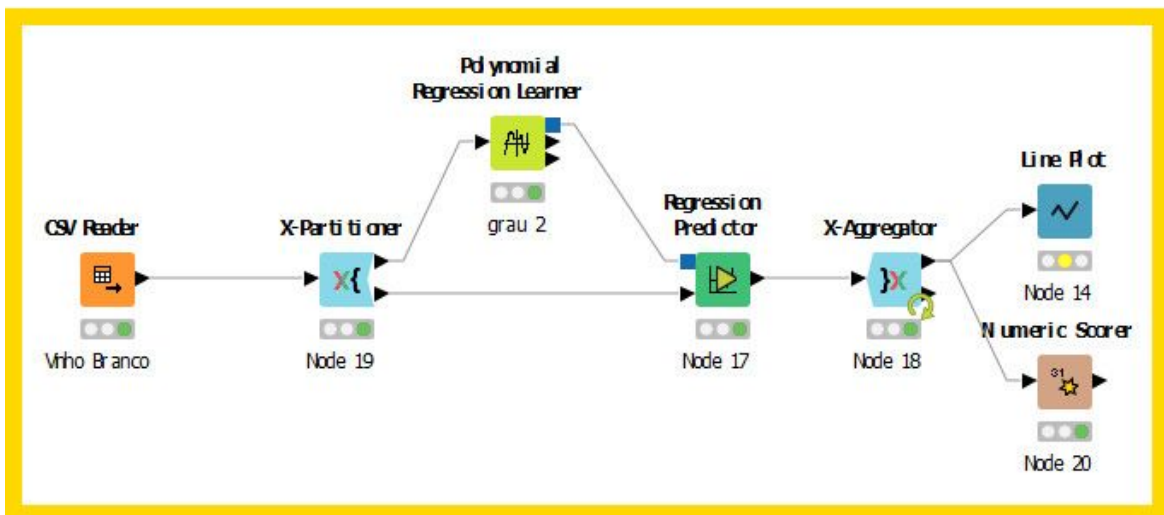
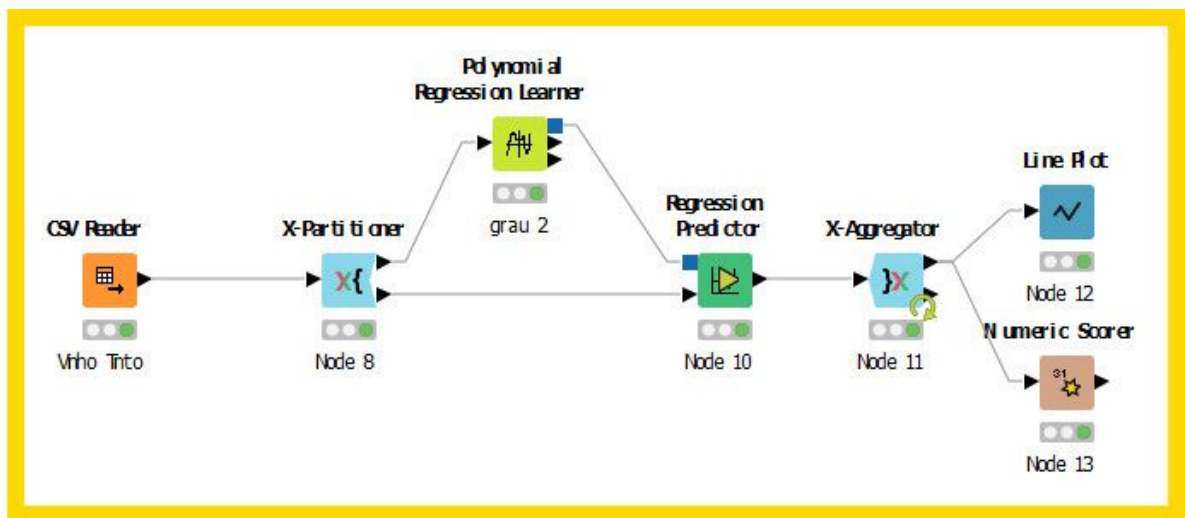
A tabela abaixo apresenta os índices do coeficiente de determinação (R^2) ou da acurácia em cada um dos conjuntos de dados de acordo com a técnica escolhida.

Técnica	Vinhos Tintos		Vinhos Brancos	
	R^2	Acurácia	R^2	Acurácia
Regressão Polinomial (grau 2)	0,366	-	0,271	-
Naive Bayes	-	57,223%	-	49,878%
Redes Neurais	-	58,974%	-	53,307%
SVM (RBF)	-	63,727%	-	57,882%

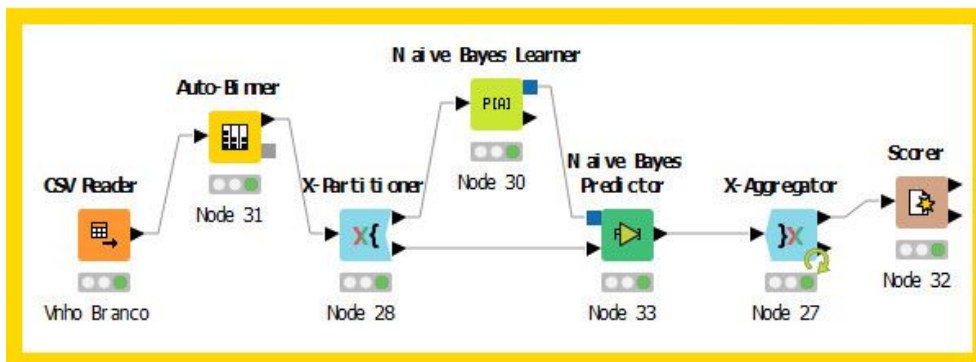
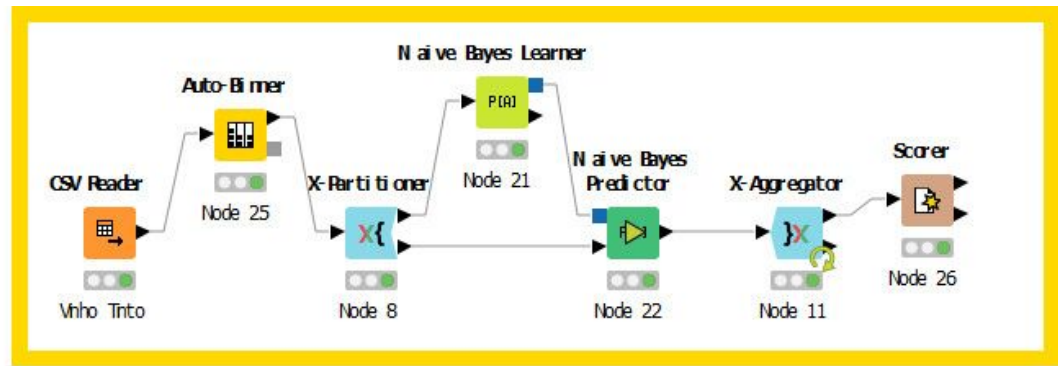
Por meio da exploração de vários modelos foi possível testar algumas hipóteses e configuração que resultaram na técnica SVM apresentando a maior acurácia de predição.

A técnica SVM apesar de ter sido inicialmente destinada a classificação binária, acaba obtendo resultados efetivos nas classificações múltiplas.

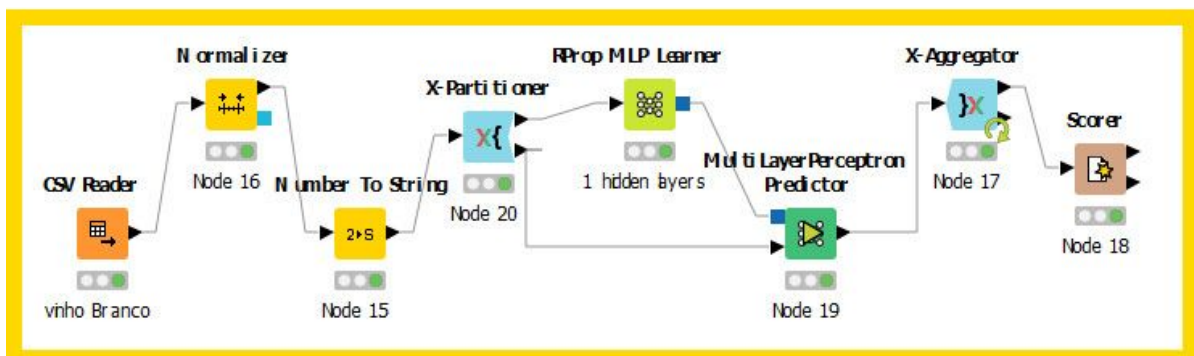
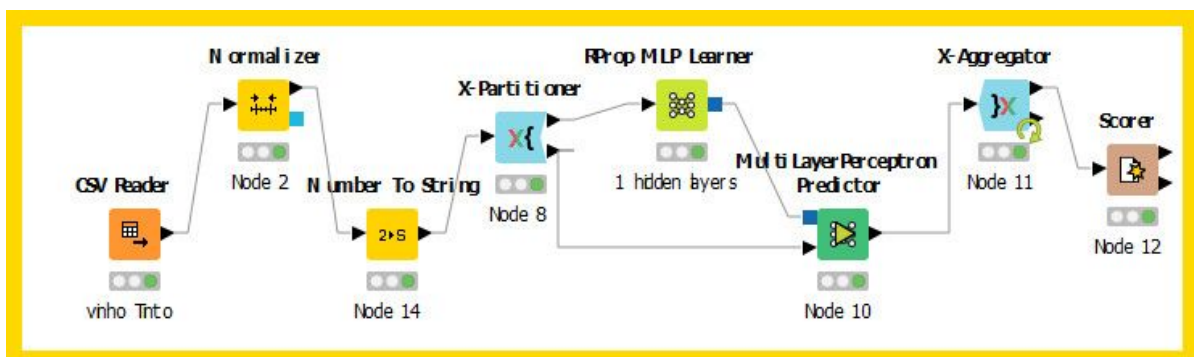
Como o propósito desta seção era de explorar a criação de modelos de machine learning para os conjuntos de dados em questão, ainda caberiam maiores testes, parametrizações e ajustes nos modelos, mas estes acabaram não sendo contemplados no presente trabalho, por restrições de prazo e escopo. No trabalho da referência [1] são citados índices de acurácia mais elevados quando é estabelecida uma margem de tolerância ($T = 1.0$) entre o resultado esperado e o predito.



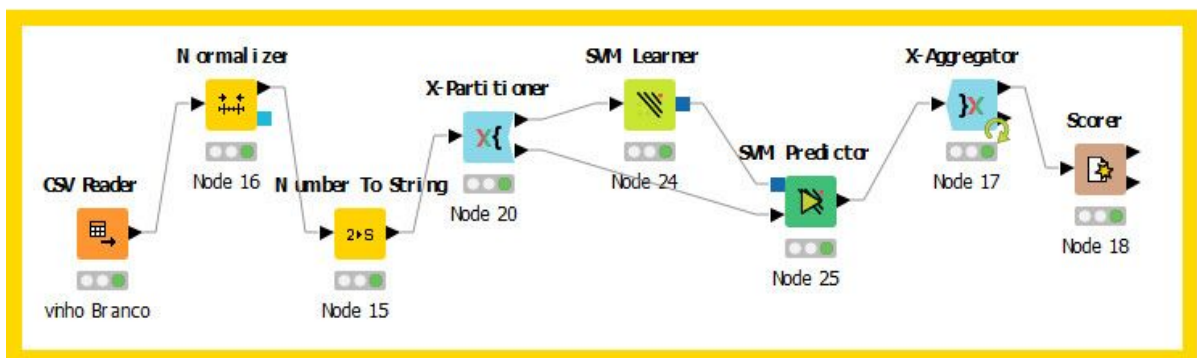
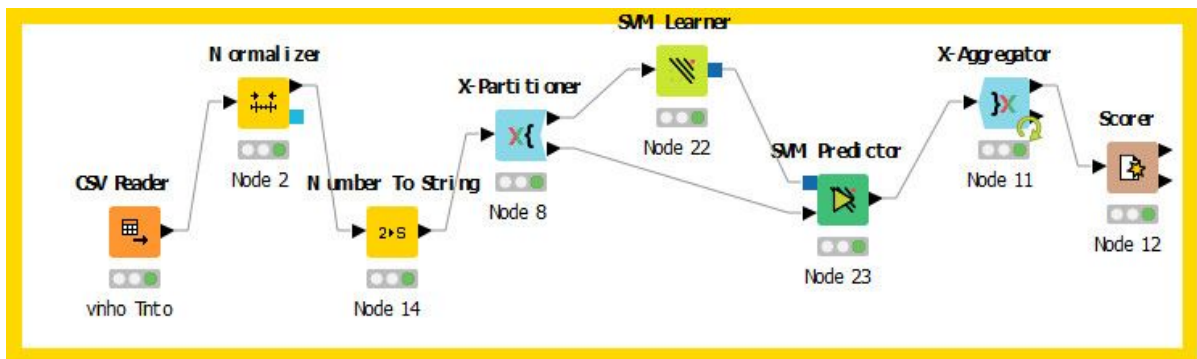
Modelo de ML com a técnica regressão polinomial na ferramenta knime para cada um dos datasets



Modelo de ML com a técnica naive bayes na ferramenta knime para cada um dos datasets



Modelo de ML com a técnica redes neurais na ferramenta knime para cada um dos datasets



Modelo de ML com a técnica SVM na ferramenta knime para cada um dos datasets

6. Apresentação dos Resultados

Utilizando o modelo de Canvas proposto por Vasandani podemos compreender alguns aspectos fundamentais do presente trabalho.

Data Science Workflow Canvas *

Título: Preditor de qualidade em vinhos		
1 Problem Statement <i>What problems are you trying to solve? What larger issues do the problem address?</i>	2 Outcomes/Predictions <i>What predictions are you trying to make? Identify applicable predictor (x) and/or target (y) variables.</i>	3 Data Acquisition <i>Where are you sourcing your data from? Is there enough data? Can you work with it?</i>
Estamos explorando um conjunto de dados relacionado com as propriedades químicas de amostras de vinhos e tentando identificar quais são as características químicas que mais influenciam a nota de qualidade que um vinho irá receber e se é possível inferir qual avaliação de qualidade um vinho terá a partir de seus atributos físico-químicos.	Inicialmente identificaremos quais das variáveis possuem maior influência na nota de qualidade que é recebida e se essas várias são as mesmas para vinhos tintos e brancos.	Utilizaremos um dataset disponibilizado no repositório de Machine Learning da UCI, que conta com mais de 5.000 amostras de vinho verde, produzido em Portugal.
4 Modeling <i>What models are appropriate to use given your outcomes</i>	5 Model evaluation <i>How can you evaluate your model's performance?</i>	6 Data preparation <i>What do you need to do to your data in order to run your model and achieve your outcomes?</i>
Exploraremos a criação de modelos voltados para os problemas de classificação, já que cada nota de qualidade pode ser interpretada como uma classe distinta.	Para avaliar a performance dos modelos compararemos os índices de acurácia obtida em cada um deles, fazendo uso da ferramenta Knime.	Como o dataset já havia sido utilizado em outros estudos acadêmicos não precisamos lidar com dados ausentes ou duplicados, ou mesmo com a descaracterização dos dados. Para alguns modelos criados foi necessária a normalização dos dados e outras parametrizações a fim de obter os melhores resultados.
* Adaptado de: https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0		

Data Science Workflow Canvas com o título "Preditor de qualidade em vinhos"

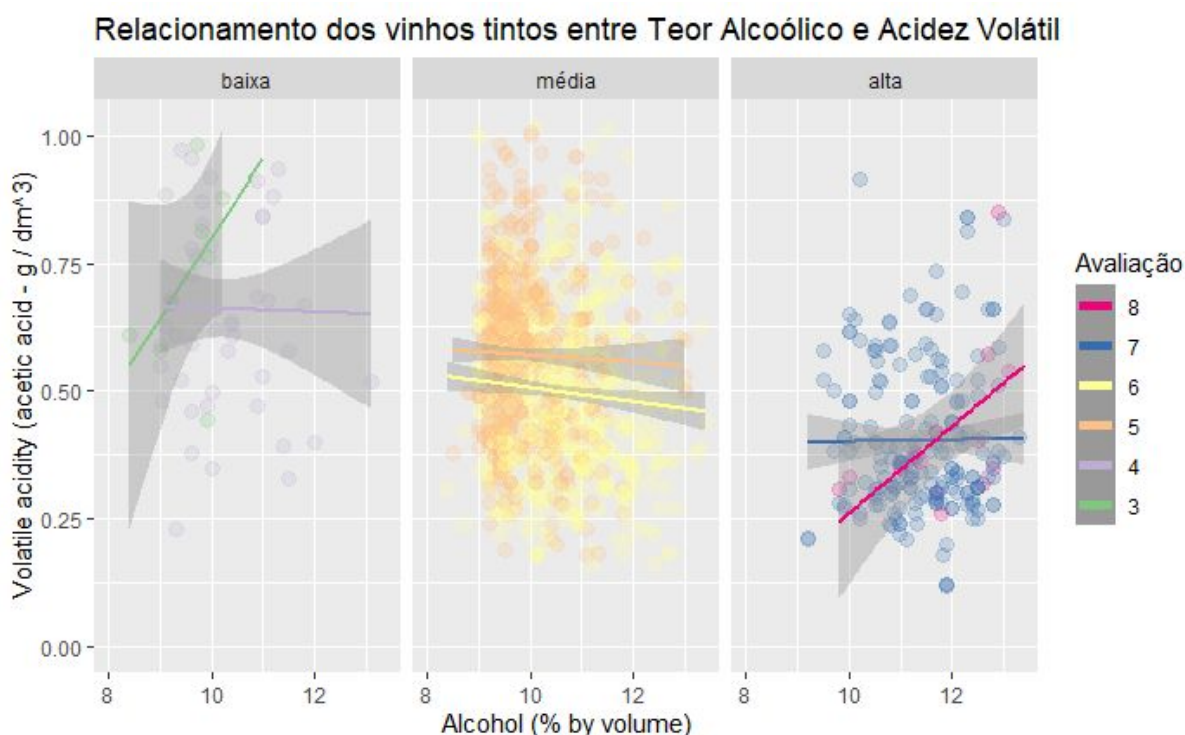
Os conjuntos de dados contemplam 6497 observações e 13 variáveis, onde a maioria delas está relacionada às propriedades químicas dos vinho. Nossa intenção com estes conjuntos de dados é investigar quais propriedades químicas influenciam a percepção de qualidade (nota) atribuída a um vinho e verificar se é possível inferir qual nota um dado vinho receberá.

Começamos conhecendo os conjuntos de dados e observando o comportamento de cada uma das variáveis e ao elaborarmos a matriz de correlação foi possível identificar as variáveis que têm maior relação com a nota de qualidade atribuída aos vinhos, o que permitiu a condução de investigações mais aprofundadas.

Pelas análises realizadas foi possível constatar que o teor alcoólico e a acidez influenciam na avaliação de qualidade que um vinho tinto receberá, mas outros fatores também podem refletir nesta avaliação, pois alguns vinhos com as notas mais altas possuem teor alcoólico e a acidez próximos aos vinhos de avaliação

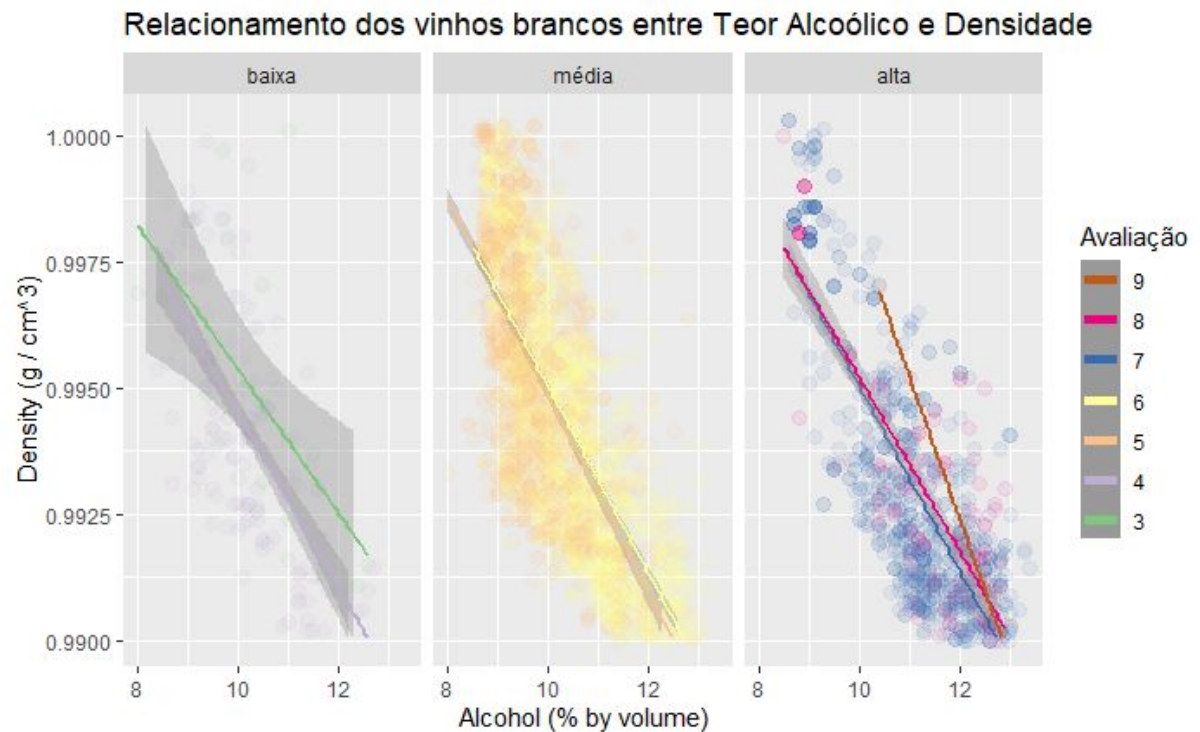
média. Já no caso dos vinhos brancos, as variáveis de maior influência foram o teor alcoólico e a densidade. E também observamos que outros fatores afetam a nota de qualidade, pois a correlação destas duas variáveis com a nota de qualidade apesar de relevante, não é muito elevada e em alguns gráficos percebemos a sobreposição entre amostras com notas de qualidade distintas.

Como os maiores índices de correlação da variável quality são com as variáveis alcohol e volatile.acidity para o conjunto de dados dos vinhos tintos, iremos representá-las em um gráfico segmentado pelo grau de avaliação, onde percebemos que existem duas concentrações mais perceptíveis, com os vinhos de avaliação mediana tendo maior acidez volátil e menor teor alcoólico e os vinhos de avaliação alta estão concentrados em uma área oposta a essa, ou seja, de menor acidez volátil e maior teor alcoólico.



Distribuição das amostras de vinho tinto, segmentadas pela avaliação (baixa, média ou alta), correlacionando o teor alcoólico com a acidez volátil

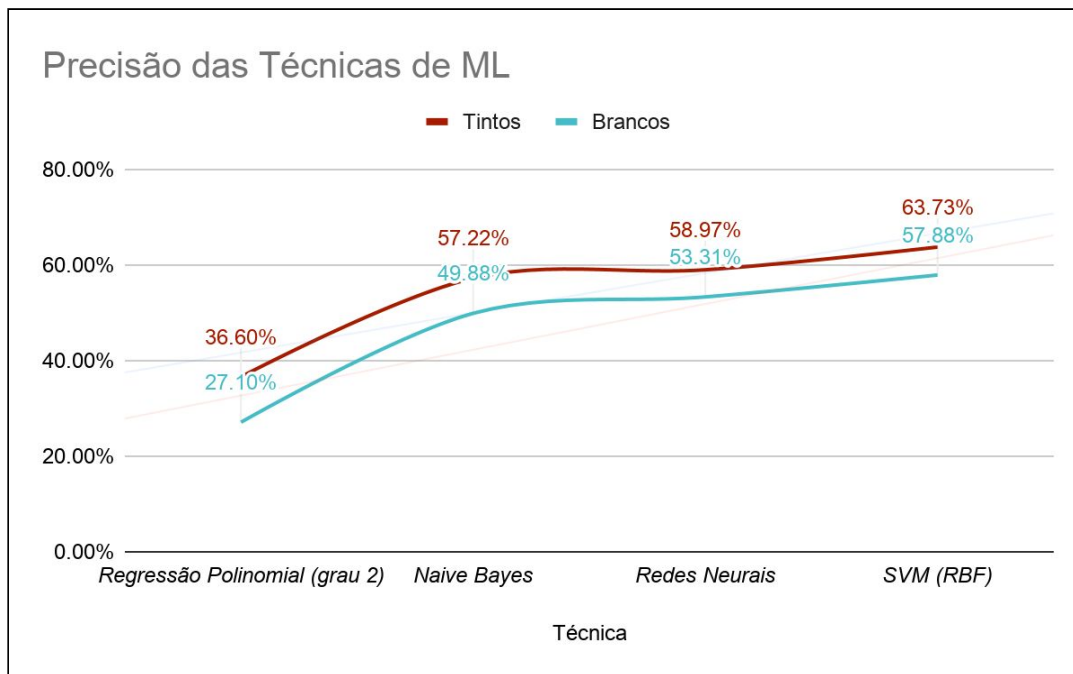
Para o conjunto de dados dos vinhos brancos temos os maiores índices de correlação da variável quality com as variáveis alcohol e density. E iremos representá-las em um gráfico segmentado pelo grau de avaliação, onde apesar de haverem sobreposições é possível identificar uma concentração dos vinhos com nota mais elevada na área que representa menor densidade e maior teor alcoólico.



Distribuição das amostras de vinho branco, segmentadas pela avaliação (baixa, média ou alta), correlacionando o teor alcoólico com a densidade

Quando utilizamos técnicas de machine learning observamos que é possível inferir a nota de qualidade que um vinho terá, a partir das características químicas de cada amostra. A técnica que obteve o maior índice de precisão foi a SVM - support vector machines fazendo uso da função de base radial - RBF para o kernel, em ambos os conjuntos de dados.

No gráfico abaixo podemos observar os índices de precisão obtidos com cada uma das técnicas utilizadas para cada um dos conjuntos de dados:



Comparação entre as técnicas de ML e a precisão da predição obtida em cada um dos conjuntos de dados

Dessa forma foi possível observar que técnicas de EDA - Análise de Dados Exploratória e ML - Aprendizado de Máquina contribuíram para o entendimento e a exploração da questão central do problema, bem como foi possível exercitar a criação de modelos que predizem a nota de qualidade que um vinho pode obter. Trabalhos futuros poderão investigar a criação e parametrização de modelos que atinjam índices de precisão mais elevados, ou que sejam capazes de sugerir ações para alavancar a nota de qualidade que um vinho poderá receber.

A referência [1] cita que as técnicas exploradas aqui podem ser utilizadas para a melhoria dos processos de avaliação e certificação de vinhos, que pode fazer parte de sistema de apoio a decisão que traria velocidade e qualidade a atuação do enólogo, ou mesmo serviria de apoio do treinamento de novos enólogos. Outro possível uso seria no controle de produção de vinhos, ajustando determinadas características como: o açúcar residual, a fermentação, o controle das bactérias ácido lácticas, etc. Há também a possibilidade de utilizar os modelos na política de preços aplicada e na identificação de públicos-alvo para uma amostra de vinho, a partir de outras amostras que tenham características similares.

7. Links

- Repositório do trabalho: <https://github.com/Erickgra/Tcc-pucminas/>
- Vídeo com a apresentação do trabalho:
<https://github.com/Erickgra/Tcc-pucminas/tree/master/video>
- TCC - versão de impressão:
<https://github.com/Erickgra/Tcc-pucminas/tree/master/TCC-versaoimpressao>
- Data Science Workflow Canvas elaborado:
<https://github.com/Erickgra/Tcc-pucminas/tree/master/DSW%20Canvas>
- Scripts de EDA: <https://github.com/Erickgra/Tcc-pucminas/tree/master/EDA>
- Modelos de ML (Knime):
<https://github.com/Erickgra/Tcc-pucminas/upload/master/arquivos-knime>

REFERÊNCIAS

- [1.] CORTEZ, Paulo et al. **Modeling wine preferences by data mining from physicochemical properties**. Guimarães - Portugal. Elsevier. 2009.
Disponível em:
<<https://repositorium.sdum.uminho.pt/bitstream/1822/10029/1/wine5.pdf>>
- [2.] VASANDANI, Jasmine. **A Data Science Workflow Canvas to Kickstart Your Projects**. 2019. Disponível em:
<<https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0>>
- [3.] CORTEZ, Paulo et al. **Wine Quality Data Set**. Guimarães - Portugal. 2009.
Disponível em: < <https://archive.ics.uci.edu/ml/datasets/Wine%2BQuality>>
- [4.] CHIH-WEI, Hsu e CHIH-JEN, Lin. **A Comparison of Methods for Multi-class Support Vector Machines**. National Taiwan University, Taipei - Taiwan.
Disponível em: <<https://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.pdf>>