Análise da explicabilidade qualitativa do vinho português

Universidade do Minho - Departamento de Informática

MIEI/MEI/MMC
Aprendizagem Automática I

Vinho & Dados

- O vinho é de importância relevante na sociedade Portuguesa.
- Perceber os fatores que formam um vinho de qualidade.
- Foi utilizado um conjunto de dados com 4898 amostras de vinho branco e 1599 amostras de vinho tinto.
- Ambos os data sets possuem 12 atributos de cariz técnico e uma classe de interesse, a qualidade, que provém de uma avaliação por júris de vinhos profissionais.

Dados e Estatísticas

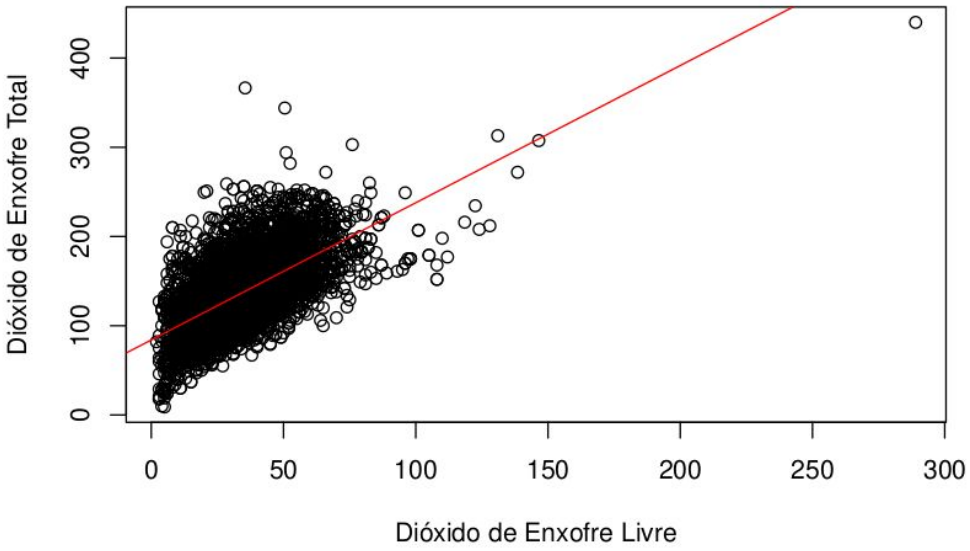
Compreensão de Negócio

- O conhecimento do significado de cada atributo é crucial para determinar a sua importância e dotar o sistema de uma base em conhecimento de domínio.
- Por este método, conseguimos perceber as, por exemplo, seguintes informações valiosas sobre o negócio:
    - Existe uma relação inversamente proporcional e entre acidez volátil e fixa.
    - Em portugal, a acidez volátil está limitada a 1.2g/L.
    - Ácido cítrico é um conservante natural, dota o vinho de frescura.
    - Açúcar residual afeta, diretamente, a doçura do vinho.
    - Dióxido de enxofre tem capacidades antioxidantes e é crucial na fermentação do vinho.

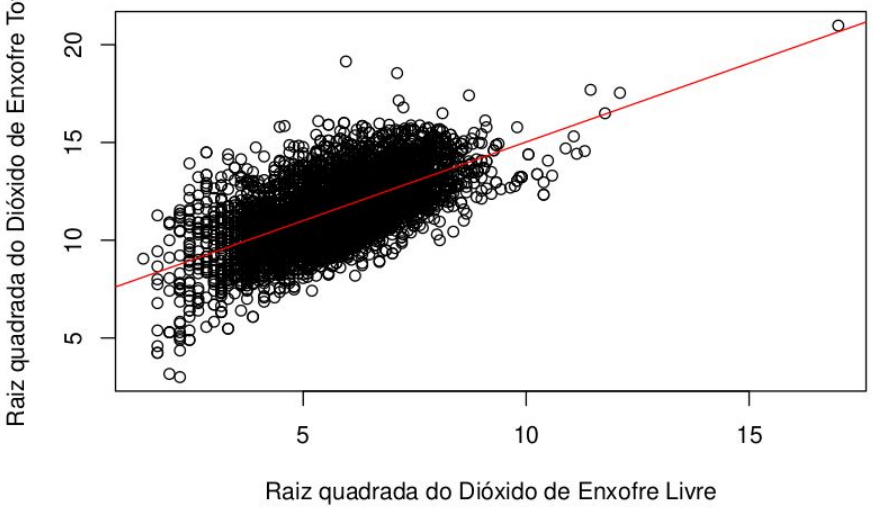
Análise Explorativa

- Não se verificou, de forma significativa, a existência de uma relação inversamente proporcional entre tipos de acidez.
- Verifica-se uma relação entre o dióxido de enxofre livre e total. No entanto, existe heterocedasticidade na relação, que oculta uma forte correlação positiva.
- Existe uma má distribuição das classes em estudo, sendo que maioria das classificações estão mais centradas para pontuações positivas.

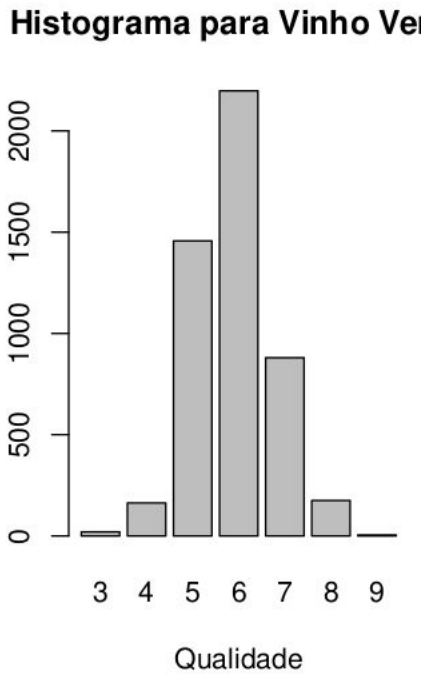
Análise Explorativa : Heterocedasticidade

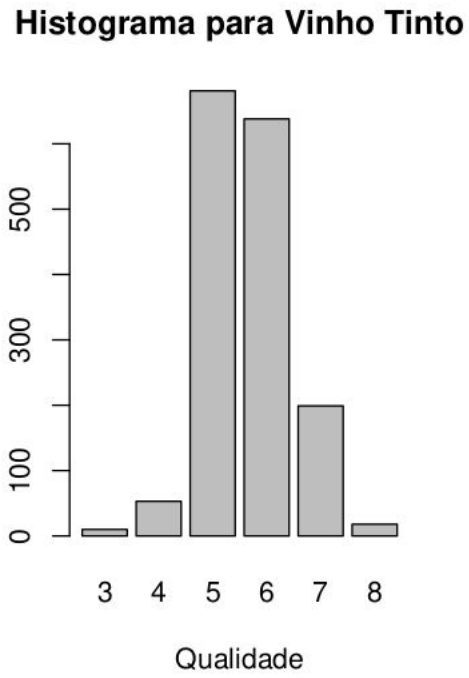


Raiz da acidez



Análise Explorativa : Qualidade





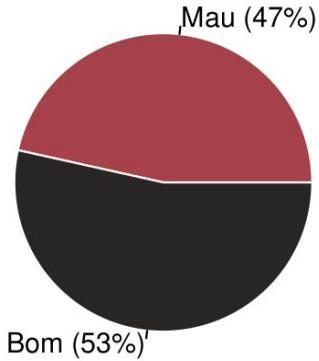
Questões Relevantes

- Quais atributos devem ser considerados para melhorar a qualidade de um vinho?
- Como lidar com a má separação de classes inerente a avaliações sensoriais?
- Qual o impacto de atributos não significativos no mercado do vinho?

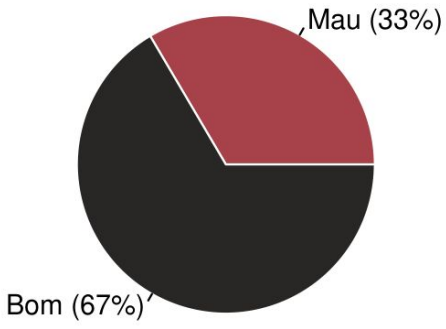
Modelação dos Dados

Regressão Logística

- Regressão logística estende o conceito de regressão para problemas de classificação.
- Este modelo é melhor adaptado a problemas de classificação com 2 classes. Tenta modelar as odds associadas às classes, que, numa relação, permite indicar a certeza das classes previstas.
- Para adaptar ao problema, as classes foram separadas em bom e mau.



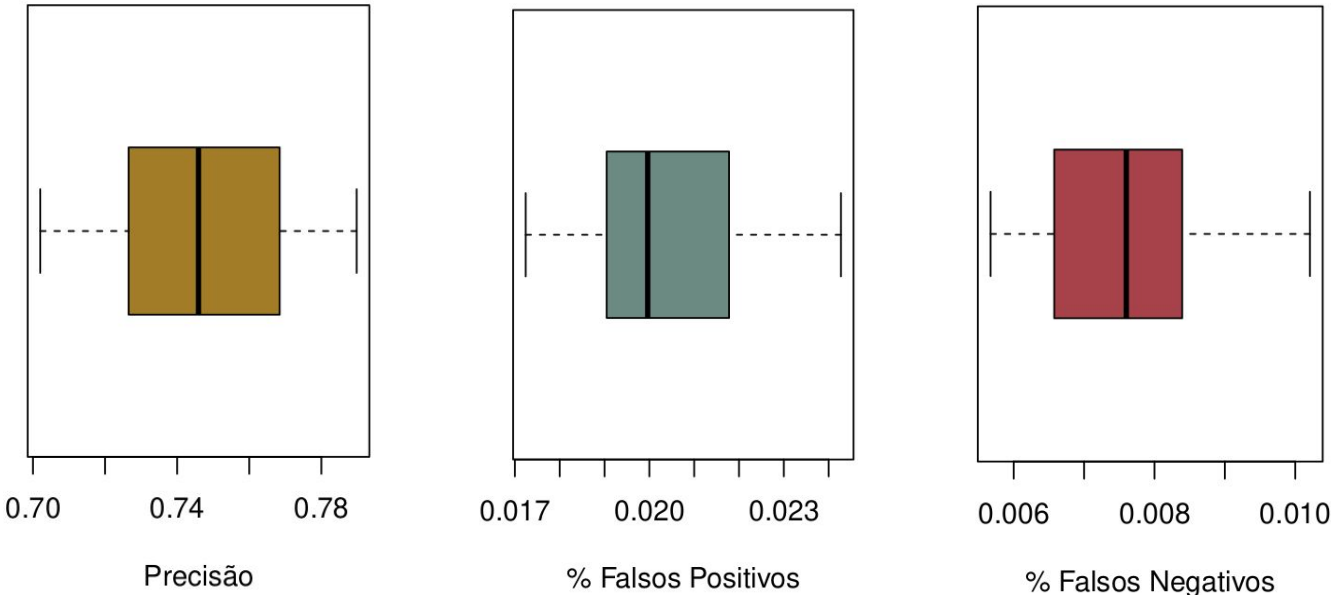
VINHO TINTO



VINHO BRANCO

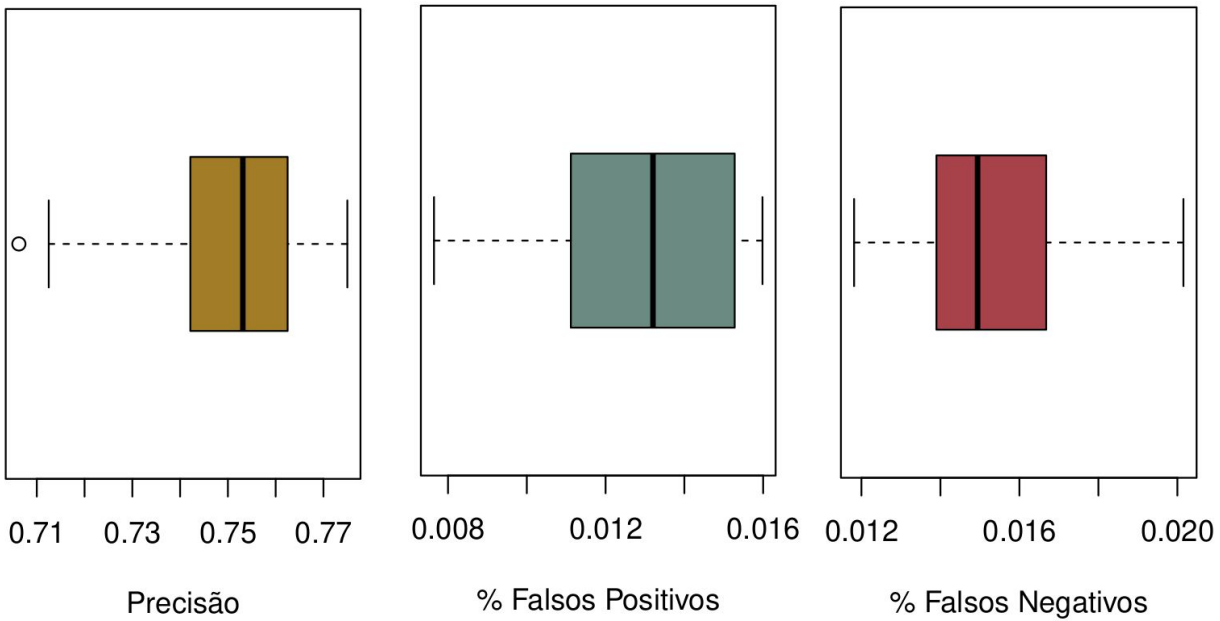
Regressão Logística : Vinho Branco

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \cdot fixed.acidity + \beta_2 \cdot volatile.acidity + \beta_3 \cdot residual.sugar + \beta_4 \cdot free.sulfur.dioxide + \beta_5 \cdot density$$

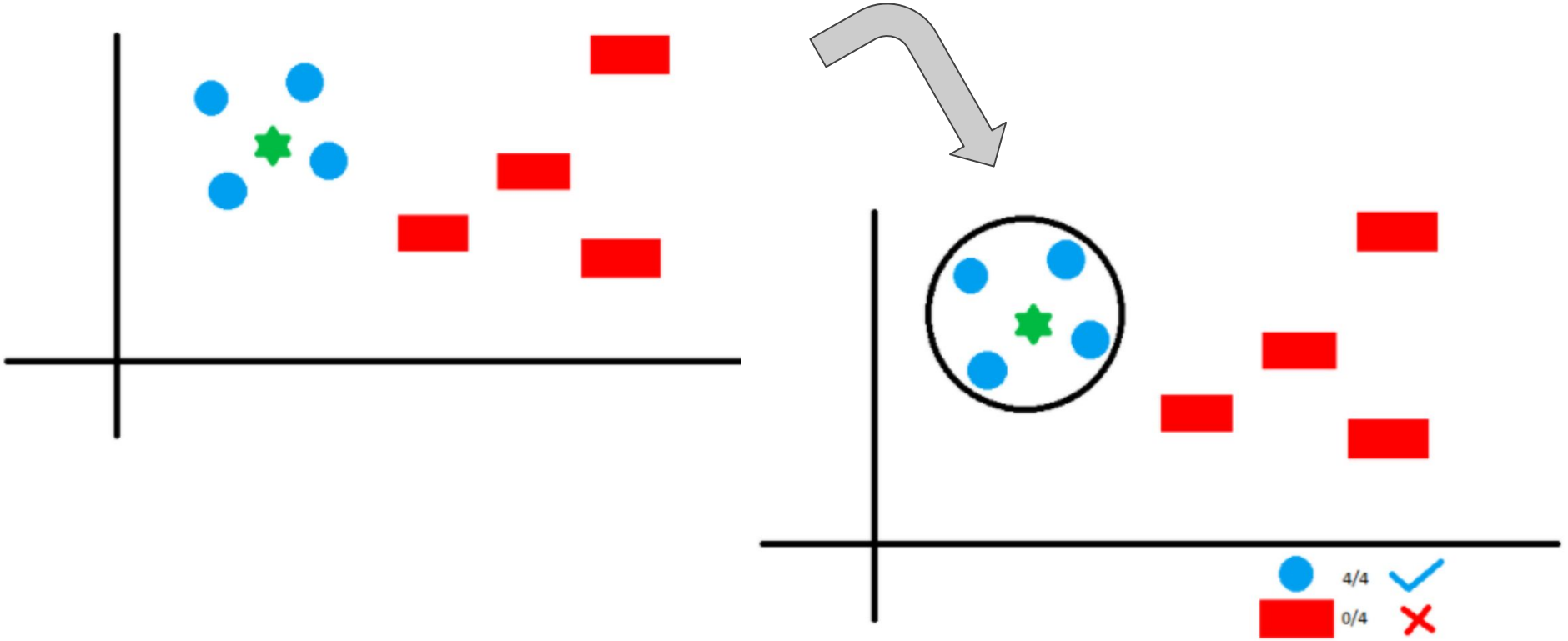


Regressão Logística : Vinho Tinto

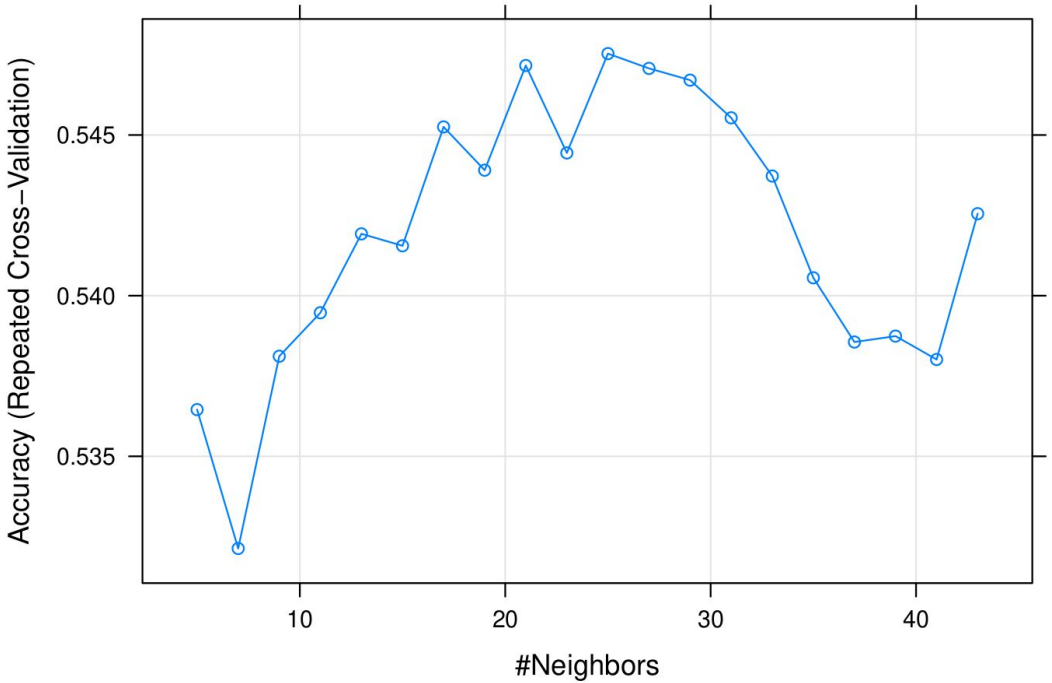
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \cdot volatile.acidity + \beta_2 \cdot chlorides + \beta_3 \cdot total.sulfur.dioxide + \beta_4 \cdot free.sulfur.dioxide$$



k-Nearest Neighbours

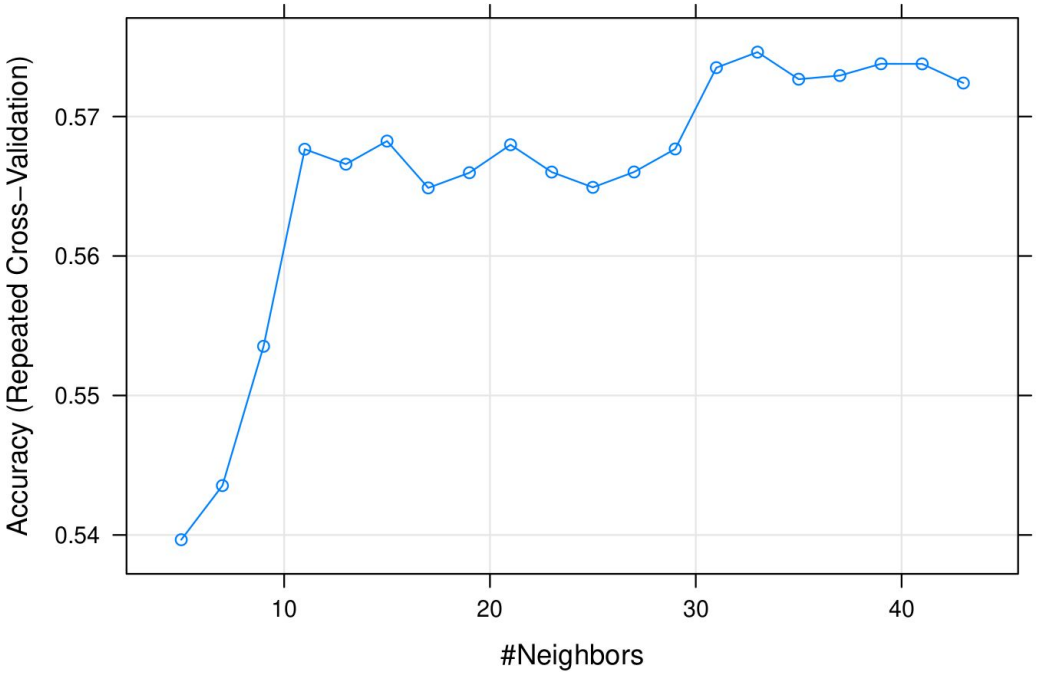


k-Nearest Neighbours : Vinho Branco




```
                    Accuracy : 0.5524
                      95% CI : (0.524, 0.5805)
No Information Rate : 0.4493
P-Value [Acc > NIR] : 3.167e-13
```

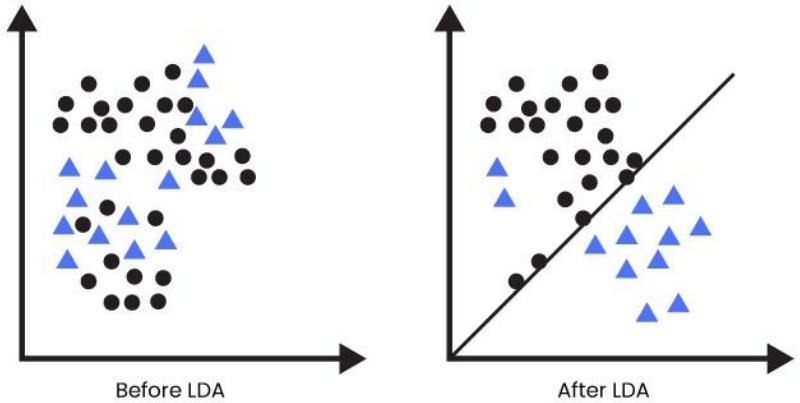
k-Nearest Neighbours : Vinho Tinto




```
                 Accuracy : 0.5844
                   95% CI : (0.5342, 0.6333)
No Information Rate : 0.4282
P-Value [Acc > NIR] : 2.896e-10
```

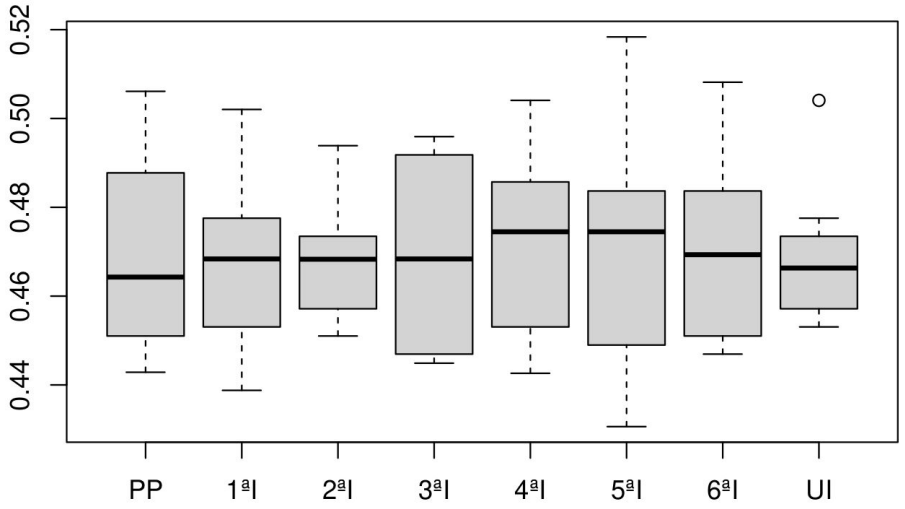
Linear Discriminant Analysis

- Permite encontrar uma combinação linear que melhor separa as classes.
- Assume que cada atributo tem a sua própria média. No entanto, as classes tem uma matriz de covariância em comum.
- O “Linear” provém da separação originada, que é linear.



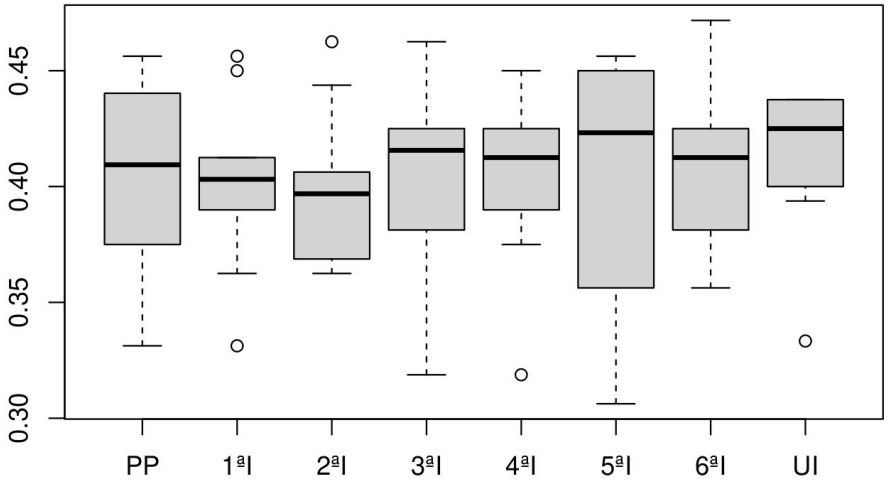
LDA : Vinho Branco

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.46917530946805
1ª Iteração	density	0.467134493141519
2ª Iteração	fixed acidity	0.467947474071596
3ª Iteração	citric acid	0.468568082970893
4ª Iteração	pH	0.469976580796253
5ª Iteração	total sulfur dioxide	0.470589662094346
6ª Iteração	free sulfur dioxide	0.469987453997993
Última Iteração	chlorides	0.468350618936099



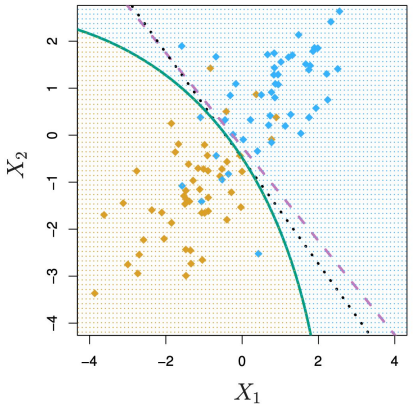
LDA : Vinho Tinto

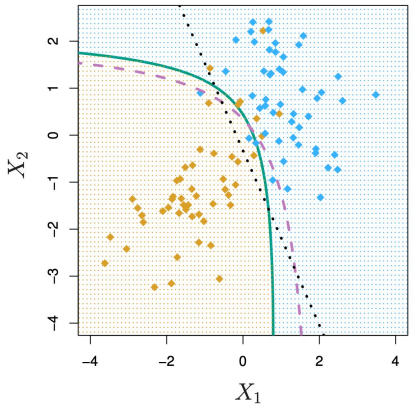
	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.404650157232704
1ª Iteração	fixed acidity	0.401493710691824
2ª Iteração	residual sugar	0.399626572327044
3ª Iteração	pH	0.409005503144654
4ª Iteração	density	0.403368710691824
5ª Iteração	citric acid	0.406517295597484
6ª Iteração	free sulfur dioxide	0.410294811320755
Última Iteração	chlorides	0.412083333333333



Quadratic Discriminant Analysis

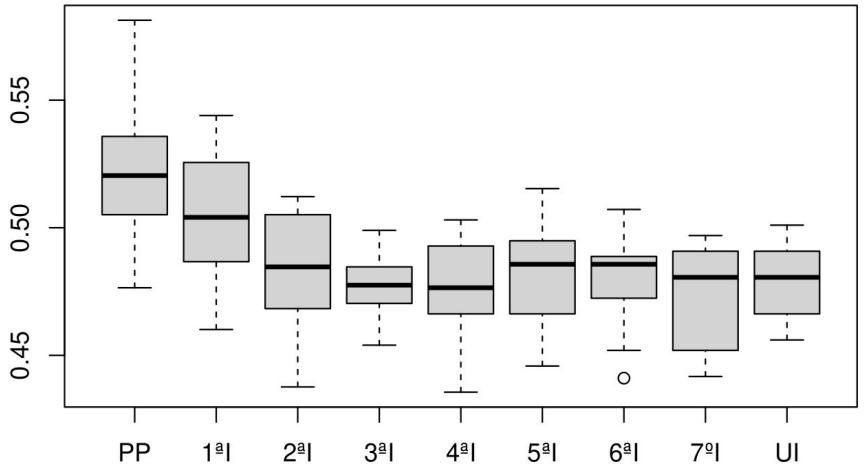
- Permite encontrar uma combinação quadrática que melhor separa as classes.
- Assume que cada atributo tem a sua própria média. Para além disso, cada classe tem uma matriz de covariância característica.
- O “Quadratic” provém da separação originada, que assemelha um polinômio de 2o grau..





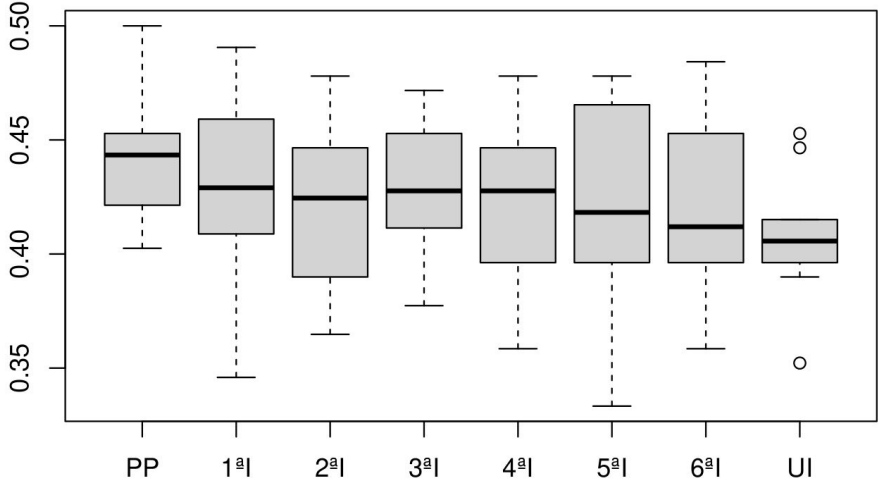
QDA : Vinho Branco

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.521320265349893
1ª Iteração	chlorides	0.504399221906329
2ª Iteração	alcohol	0.482712354730909
3ª Iteração	citric acid	0.478232081400569
4ª Iteração	free sulfur dioxide	0.47802134769814
5ª Iteração	total sulfur dioxide	0.480298269240361
6ª Iteração	sulphates	0.47927951518779
7ª Iteração	fixed acidity	0.474367798892713
Última Iteração	pH	0.479878048780488



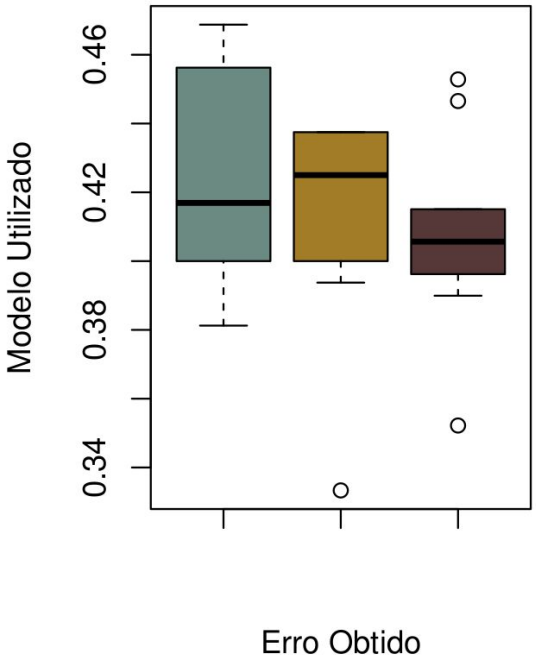
QDA : Vinho Tinto

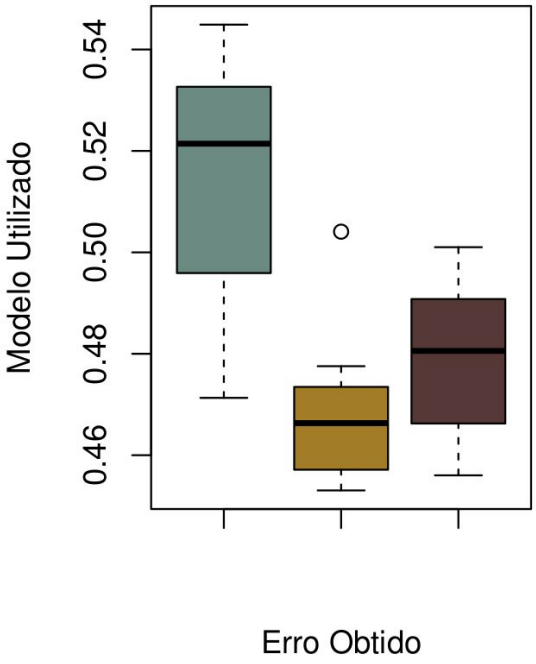
	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.441194968553459
1ª Iteração	pH	0.427314704243293
2ª Iteração	chlorides	0.420412387548762
3ª Iteração	free sulfur dioxide	0.429189554971738
4ª Iteração	density	0.422307141151182
5ª Iteração	fixed acidity	0.421634424010827
6ª Iteração	citric acid	0.417888703128732
Última Iteração	residual sugar	0.407797946023406



Conclusões

Comparação entre Modelos





Questão nº1 - Quais atributos devem ser considerados para melhorar a qualidade de um vinho?

- Nos vinhos brancos, a medida de qualidade está significativamente relacionada com a tipicidade do vinho: se é verde, maduro, ou do douro.
- No entanto, nos vinhos tintos, a qualidade recai mais sobre a qualidade da fermentação do mesmo, o que é evidenciado pela significatividade de atributos relacionados com a fermentação do vinho. Nomeadamente, o valor total de dióxido de enxofre.

Questão nº2 - Como lidar com a má separação de classes inerente a avaliações sensoriais?

- As classificações médias dos júris não modeladas, o que aponta para um erro irredutível associado a este parâmetro.
- Porém, com a informação existente é impossível desenvolver modelos mais complexos.
- Nesse sentido, separar as classes em duas, “vinho mau” e “vinho bom”, fornece um modelo primitivo para reduzir o impacto do bias associado à classificação de cada júri.

Questão nº3 - Qual o impacto de atributos não significativos no mercado do vinho?

- Conhecer os atributos não significativos permite dotar uma empresa de vinhos das ferramentas para modelar um vinho com o intuito de maximizar a apreciação de mercado.
- No entanto, há características do vinho que, apesar de não relevantes na medida de qualidade, são absolutamente necessárias para a criação do vino em si.
- Por esta razão, este conhecimento deve servir de ferramenta de auxílio a profissionais, e nunca como uma receita para obter um vinho de qualidade.

Realizado por:

- Nelson Estevão, a76434 - MIEI
- Rui Reis, a84930 - MIEI
- Susana Marques, a84167 - MIEI

Análise da explicabilidade qualitativa do vinho português

Universidade do Minho - Departamento de Informática

MIEI/MEI/MMC
Aprendizagem Automática I