Guiao Apresentacao TP AA1

# Análise Exploratória

## José Pedro

Realizamos a análise exploratória para perceber logo à partida qual a influência de cada preditor tanto na variável qualitativa, como na variável quantitativa. Isto era importante para tentar eliminar alguns preditores à partida, visto que estávamos a lidar com muitos e era necessário perceber quais os mais importantes.

Para além disto, verificamos quais preditores nos eliminavam muitos casos, devido aos NAs (bp.2s e d) e eliminamos esses preditores visto reduzirem o dataset para perto dos 25%.

Verificamos também as correlações entre as diferentes variáveis para tentar perceber quais se relacionam. Isto é importante para não se incluir no modelo 2 variáveis que estão altamente relacionadas, pois uma já explica a outra e não traz grande coisa de novo ter mais uma, apenas complexidade.

Para além disto, e visto que alguns métodos estudados (com qda e lda) usam pressupostos da normalidade das variáveis decidimos verificar quais as variáveis que se aproximam de seguir uma distribuição normal, para eliminar logo à partida essas, no uso desses métodos.

## Carlos

Fala da análise à influência

Colocar gráficos que achar necessários

## Ricardo

Fala da análise à normalidade

Colocar gráficos que achar necessários

# Seleção do Modelo

## José Pedro

Explica que dividimos o trabalho pelos 3, ficando:

* Ricardo -> regressão linear (variável quantitativa) e posterior classificação
* Carlos -> regressão logística
* José Pedro -> lda, qda e knn

Explica que inicialmente decidimos utilizar um método no qual treinávamos o modelo com o dataset inteiro e o testávamos também com o mesmo, no entanto para estes casos estaríamos a beneficiar os modelos com overfitting.

Depois passamos para utilizar uns dados de teste e outros de treino, mas aleatórios. No entanto, este método é bastante instável, pelo que passamos a utilizar validação cruzada com k-folder (k=10) com o dataset total. A ideia passou por cada um selecionar um (ou dois) modelos que achasse melhor e depois comparamos cada um dos modelos selecionados por cada um. A comparação entre os “melhores modelos” já utilizou a base de dados dividida em casos de teste e treino.

Para as comparações entre modelos, não utilizamos apenas o acerto como sendo o fator de decisão, mas também a complexibilidade (+baixa->melhor) e interpretabilidade (+alta->melhor) do mesmo.

CADA UM FALA DAS SUAS ABORDAGENS PELA SEGUINTE ORDEM E POR IMAGENS RELEVANTES

1. José Pedro
2. Carlos
3. Ricardo

## Carlos GLM

Para tentar perceber qual o melhor threshold, e visto que nem sempre o melhor é 0.5, decidimos criar uma função que vai variando o threshold (consoante os argumentos passados) e que retorna o acerto para cada threshold. Os valores de acerto retornados, não são simplesmente o acerto total mas:

* Acerto total
* Acerto nos casos positivos
* Acerto nos casos negativos

Porque não é só importante o acerto total, mas também não obter muitos falsos positivos, nem negativos.

Dizer que realizamos alguns gráficos com estes valores, para perceber qual o melhor valor (melhor tradeoff) entre os 3.

MOSTRAR GRÁFICO

Explicar a função que escolhemos, como sendo uma função de um acerto ponderado na qual damos o dobro da ponderação para o acerto total, relativamente aos restantes, mas no qual contabilizamos os restantes. Explicar que podíamos ter seguido outras abordagens e que pensamos em algumas, como por exemplo um acerto abaixo de um X em um dos 3 colocava o acerto a 0, mas que achamos que esta é equilibrada o suficiente.

# Seleção do melhor modelo e resposta às perguntas

O melhor modelo selecionado foi o de regressão linear com a seguinte expressão:

Glyhb = 1.693028 + 0.027309 \* stab.glu + 0.019922\* age

E depois classificamos consoante o valor da glyhb, da seguinte forma:

* Superior a 7, então tem diabetes
* Caso contrário, não tem

Depois falar nos resultados.