14 DE NOVEMBRO DE 2018

Universidade do Minho
Escola de Engenharia

# APRENDIZAGEM AUTOMÁTICA I
# TRABALHO PRÁTICO

## PARTE 1

Carlos José Lima Gonçalves, a77278
José Pedro dos Santos Ferreira, a78452
Ricardo Jorge Marques Peixoto, a 78587

***Dataset*:** diabetes

O problema consiste em prever se um determinado indivíduo contém ou não diabetes, consoante alguns fatores (neste caso apenas foram observadas pessoas Afro-Americanas que residem no centro do estado de *Virginia*).

No conjunto de dados existem 403 testes e 19 variáveis. As variáveis são as seguintes:

- **id:** ID da pessoa (não é para considerar)
- **chol:** colesterol total
- **stab.glu:** glucose estabilizada
- **hdl:** Lipoproteína de alta densidade
- **ratio:** rácio entre colesterol/hdl
- **glyhb:** hemoglobina glicada
- **location:** cidade (um fator com opções Buckingham e Louisa)
- **age:** idade em anos
- **gender:** género (fator com opções masculino e feminino)
- **height:** altura em polegadas
- **weight:** peso em libras
- **frame:** relação entre a circunferência do pulso relativamente à altura (um fator com opções pequeno, médio e grande)
- **bp.1s:** 1º pressão arterial sistólica (quando o coração bombeia sangue)
- **bp.1d:** 1º pressão arterial diastólica (quando o coração não está a bombear o sangue)
- **bp.2s:** 2º pressão arterial sistólica
- **bp.2d:** 2º pressão arterial sistólica
- **waist:** cintura em polegadas
- **hip:** anca em polegadas
- **time.ppn:** tempo, em minutos, após a refeição

A variável resposta depende do nível de hemoglobina glicada, visto que o seu valor acima de 7 é geralmente considerado um teste positivo de diabetes.
Ao observarmos estes dados vemos que se trata de um problema supervisionado, de classificação.
Podemos realizar este trabalho utilizando técnicas de classificação nas quais obtemos a probabilidade de ter diabetes, no entanto e visto que conhecemos o método de classificação de ter ou não diabetes, podemos também usar um método de regressão para prever o valor da hemoglobina e depois classificar consoante o valor obtido.

### Questões às quais vamos tentar responder:

- Será que o colesterol/pressão arterial/tempo após refeição/fatores corporais afetam os diabetes?
- Qual fator corporal explica melhor o valor da diabete?
- Quais fatores influenciam mais o resultado final?
- De que forma os fatores selecionados para a explicação dos resultados o influenciam? (crescentemente, decrescentemente, linearmente)
- Qual a probabilidade (ou confiança no resultado) de uma pessoa com o colesterol X(aqui podíamos por caraterísticas X e depois fazíamos para várias) ter diabetes?
- Qual a taxa de incidência em pessoas com menos e com mais de 50 anos?
- Qual país apresenta maior incidência? (visto serem só dois podemos comparar)
- O resultado é mais exato utilizando um modelo de classificação ou de regressão (e de seguida classificando)?