Aprendizagem Automática
Trabalho Laboratorial – grupos de 2 ou 3 alunos
Pima Indians Diabetes Dataset

1 Dados

Os dados disponibilizados encontram-se no ficheiro pickle: `pimaDiabetes.p`. O *Pima Indians Diabetes Dataset*, originalmente desenvolvido pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais dos Estados Unidos da América, contém informações sobre 768 mulheres de uma das tribos indígenas desse país[1]. Devido a uma alta taxa de incidência de diabetes, a população dos índios Pima tem estado sobre contínua vigilância pelo Instituto desde 1965, cada residente tendo sido submetido a um exame padronizado cada dois anos. Nos dados disponibilizados, 500 mulheres não têm diabetes e 268 sim. A informação recolhida consiste nas seguintes 8 características:

Característica	Gama de valores
Número de gravidezes	[0,17]
Concentração de glicose plasmática em 2 horas num teste oral de tolerância à glicose	[0,199]
Pressão arterial diastólica (mm Hg)	[0,122]
Espessura da dobra cutânea do tríceps (mm)	[0,99]
Nível sérico de insulina em 2 horas (μh/ml)	[0,846]
Índice de massa corporal (peso em kg/altura em m)	[0,67.1]
Diabetes Pedigree Function	[0.078,2.42]
Idade (anos)	[21,81]

2 Objetivos

Em termos globais, o que se pretende é determinar se os pacientes têm diabetes baseado nos indicadores disponibilizados pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais. Neste contexto, é necessário treinar e avaliar três ou mais classificadores binários e fazer um estudo comparativos do desempenho dos modelos escolhidos.

[1] Os índios Pima, também conhecidos como Akimel O'odham (O'odham para "gente do rio") vivem no centro e sul do Arizona, bem como o noroeste do México nos estados de Sonora e Chihuahua.
Acredita-se que o nome abreviado, "Pima", tenha vindo da frase *pi'añi mac* ou *pi mac*, que significa "não sei", termo que eles usaram repetidamente nos encontros iniciais com os colonos espanhóis. Os espanhóis referiam-se a eles como Pima. Este termo foi adotado posteriormente por comerciantes, exploradores e colonos ingleses.

3 Desenvolvimento

Deverá ter em conta os seguintes pontos:

1. **Modelos de Classificação**:
   (a) Escolher 3 classificadores binários. Um dos classificadores tem de ser o "`RandomForestClassifier`".
   (b) Treinar os classificadores binários, tendo em conta a escolha dos hiper-parâmetros dos mesmos.
   (c) Escolher a metodologia de treino/teste apropriada de modo a ter uma estimação fidedigna do desempenho dos modelos treinados.
   (d) Usar as métricas apropriadas e calibrar os modelos treinados.
   (e) Fazer um estudo comparativo do desempenho do classificadores
2. **Pré-processamento dos dados**:
   Investigue se normalizar os dados[2] é benéfico para o desempenho dos classificadores.
3. **Observações Gerais**:
   Deve saber justificar as escolhas feitas no trabalho, tanto a nível da escolha das metodologias de treino/testes usadas, como na seleção dos classificadores implementados. Adicionalmente, deve também fazer uma análise rigorosa dos resultados obtidos.

4 Ficheiro a Entregar

Deve ser entregue, via Moodle, um único ficheiro Jupyter denominado `AxxxxxAxxxxxAxxxxxTP1.ipynb` (onde `Axxxxx` são os números de alunos do grupo - colocar em ordem crescente). O Jupyter Notebook deve estar devidamente comentado de modo a transmitir claramente os passo dados, a razão para as opções tomadas e uma análise detalhada dos resultados obtidos.

[2] Transformar os dados de maneira a cada dimensão ter média nula e variância unitária.