Trabalho 2

No contexto de um estudo realizado pela Columbia University foram recolhidos, entre 2002 e 2004, dados sobre participantes em eventos experimentais de *speed dating*. Os participantes foram recrutados entre os estudantes da universidade. Na experiência, a marcação de encontros foi gerada por correspondência aleatória e também foi feita uma variação aleatória no número de parceiros em potencial.

Durante a experiência, os participantes teriam um "primeiro encontro" de quatro minutos com todos os outros participantes. Ao final de seus quatro minutos, os participantes foram questionados se gostariam de ver seu par novamente. O conjunto de dados também inclui as respostas de um questionário feito aos participantes em diferentes pontos do processo.

Para este trabalho prático foi selecionado um subconjunto de atributos recolhidos ao longo desta experiência e cuja descrição é apresentada na Tabela 1.

O objetivo do trabalho prático é criar e avaliar modelos de classificação para este conjunto de dados recorrendo a dois algoritmos de *Machine Learning:* Árvores de Decisão (CART ou ID3) e Naive Bayes.

A avaliação do trabalho é composta por duas componentes, com as seguintes ponderações:

- Código usado: 30%
- Relatório técnico: 70%

Código

A implementação pode ser feita em Java ou Python. Poderão ser usadas bibliotecas específicas para *Machine Learning*, não sendo necessário implementar os algoritmos caso já existam implementações dos mesmos nessas bibliotecas.

Bibliotecas recomendadas:

- JSAT (Java):
  - Código: https://github.com/EdwardRaff/JSAT
  - Referência: https://www.edwardraff.com/jsat_docs/JSAT-0.0.8-javadoc/
  - Exemplos: https://github.com/EdwardRaff/JSAT-Examples/blob/master/README.md
- Pandas + Scikit-learn (Python):
  - Pandas (para ler e manipular dados): https://pandas.pydata.org/
  - Scikit-learn (biblioteca com algoritmos): https://sklearn.org/

Recomenda-se a utilização de Python, dada a grande quantidade de documentação online e bibliotecas de *Machine Learning* disponíveis para esta linguagem.

Relatório

O relatório deverá ter entre 10 e 15 páginas e deverá ser composto pelas seguintes secções:

- Título e autores
- Resumo (máximo 250 palavras)
- Introdução: descrição breve do problema e da metodologia, linguagem de programação e bibliotecas usadas (ver abaixo)
- Algoritmos: descrição dos algoritmos usados
- Análise exploratória dos dados: descrição do conjunto de dados, análise das variáveis e descrição dos passos de pré-processamento
- Experiências e resultados: descrição da metodologia experimental adotada, métricas usadas e resultados obtidos. Deverá conter uma secção com uma breve discussão crítica dos resultados, e referir as limitações do trabalho, bem como dificuldades na sua execução.
- Conclusões (máximo 250 palavras)
- Referências bibliográficas

Desenvolvimento do trabalho e prazos:

- **Os grupos devem ser constituídos por dois elementos e, preferencialmente, manter a constituição do primeiro trabalho prático**. A realização do trabalho individualmente poderá ser admitida mas não é encorajada. As alterações de grupo devem ser **comunicadas aos docentes**.
- O prazo limite para submissão do trabalho é **6 de Junho**.
- Devem **submeter o código e um relatório** com a descrição dos métodos implementados e análise experimental com discussão dos resultados obtidos.
- Os grupos podem discutir abordagens mas não podem partilhar código. Caso os programas incluam excertos que o grupo reutilizou, devem identificar claramente onde começa e onde termina o código que o grupo não implementou. Devem procurar implementar programas bem estruturados.
- Todas as referências utilizadas devem ser incluídas no relatório.
- A **apresentação** dos trabalhos decorrerá na **semana de 7 a 12 de Junho**.
- Os elementos do grupo podem dividir o trabalho entre si. Contudo, espera-se que qualquer um dos elementos conheça e possa descrever o trabalho efetuado por si e pelo outro elemento do grupo.
- Os exames da UC poderão ter questões relacionadas com os trabalhos práticos.

Tabela 1: Descrição dos atributos.

Atributo	**Descrição**
id	número de identificação do participante
partner	número de identificação do par
age	idade do participante
age_o	idade do par
goal	Qual é o seu objetivo principal ao participar neste evento? Passar uma noite divertida = 1 Conhecer novas pessoas = 2 Conseguir um encontro = 3 Procurar um relacionamento sério = 4 Dizer que consegui = 5 Outro = 6
date	Em geral, quão frequentemente sai para encontros? Várias vezes por semana = 1 Duas vezes por semana =2 Uma vez por semana = 3 Duas vezes por mês = 4 Uma vez por mês = 5 Várias vezes por ano = 6 Quase nunca = 7
go_out	Com que frequência sai (não necessariamente para encontros)? Várias vezes por semana = 1 Duas vezes por semana =2 Uma vez por semana = 3 Duas vezes por mês = 4 Uma vez por mês = 5 Várias vezes por ano = 6 Quase nunca = 7
int_corr	Correlação entre os *ratings* de interesses (desporto, museus, caminhadas, música, filmes, livros, etc.) do participante e do seu par ([-1,1]).
length	A duração de 4 minutos para o encontro é: Demasiado curta = 1 Demasiado longa = 2 Adequada = 3
met	Já conhecia o seu par anteriormente? (0/1)
like	Quão gostou do seu par? (escala 1-10; nada = 1; muito =10)
prob	Qual é a probabilidade do seu par ter gostado de si? (escala 1-10; pouco provável = 1; muito provável =10)
match	Há *match*? (0/1) (variável objetivo)