Universidade do Minho
Departamento de Informática

Perfil Sistemas Inteligentes
Aprendizagem e Extracção de Conhecimento
Edição 2018/2019

Trabalho prático – 2ª Fase

Tema	EXTRAÇÃO DE CONHECIMENTO
Objetivos de aprendizagem	Com a realização deste trabalho prático pretende-se que os alunos aprendam os seguintes procedimentos utilizados em Projetos de Extração de Conhecimento: • Extração e preparação de dados a serem utilizados para previsões; • Desenvolver/preparar *features* e *target variables* para o treino de modelos de Extração de Conhecimento; • Treinar modelos; • Aceder à *performance* do modelo através de dados de teste.
Enunciado	Este enunciado pretende ser o ponto de partida para o desenvolvimento de um modelo preditivo utilizando o ambiente de desenvolvimento Python/Sklearn. Para isso, será necessário o desenvolvimento de uma solução para o seguinte problema: *Preparação e análise de um dataset de comportamentos biométricos de alunos durante exames como forma de prever os níveis de stresse percecionados.* Numa sociedade cada vez mais exigente, locais de trabalho/académicos têm sido os principais ambientes onde cada vez mais casos de esgotamento mental surgem com regularidade. Desta forma, uma das necessidades para a prevenção destes casos passa pela monitorização não invasiva do estado de stresse de indivíduos durante tarefas de alto risco em plataformas informáticas. Com a progressão de novas técnicas de Extração de Conhecimento e de Aprendizagem, uma das soluções deste sistema baseia-se na análise em tempo-real do comportamento biométrico de utilizadores para prever o seu nível de stresse. Anexo a este trabalho prático encontram-se dois *datasets*, apresentando o conjunto de *features* biométricas (i.e., análise do uso do rato ou tomada de decisões) de alunos universitários, adquiridas durante exames académicos. Para cada instância/caso de estudo de cada *dataset*, além de apresentar o conjunto de features relativos ao comportamento biométrico, encontra-se associado um valor numérico PSS [0-52], representando o nível percecionado de stresse fornecido por cada aluno durante cada exame. Quanto maior o valor PSS, maior o stresse percecionado pelo aluno. Definido o problema, faz-se uma breve explicação de cada uma das *features*: ***Features* - Dinâmica do Rato:** - *Absolute Sum of Degrees* (ASD): calcula quanto o rato virou, independentemente do lado de direção que virou (medido em graus); - *Average Excess of Distance Between Clicks* (AED): excesso de distância no qual o rato "viajou" entre dois cliques do botão do rato (medido em píxeis); - *Click Duration* (CD): tempo entre o clique e o largar do botão do rato (medido em milissegundos); - *Distance Between Clicks* (DBC): distância "viajada" pelo rato entre dois cliques de rato consecutivos (medido em píxeis); - *Mouse Velocity*: distância "viajada" pelo rato durante um determinado período de tempo (medido em píxeis/milissegundo); - *Mouse Acceleration*: velocidade do rato medido sobre o tempo (medido em píxeis/milissegundo^2). - *Time Between Clicks*: tempo entre o clique e o largar do botão do rato, i.e., quanto tempo demorou um indivíduo a executar outro clique de rato (medido em milissegundos).

***Features* – Tomada de Decisão:**

- *Average Time Between Decision* (ATBD): tempo médio que cada aluno apresenta para tomar uma decisão. As decisões analisadas variam desde entrada/saída de uma questão, inserção, alteração ou remoção de uma resposta, marcação/desmarcação de questões para revisão, entre outras (medido em milissegundos);
- *Median Time Between Decision* (MTBD): mediana do tempo que cada aluno demora a tomar uma decisão (medida em milissegundos);
- *Standard Deviation/Variance Time Between Decision*: desvio padrão/variância do tempo entre as decisões para cada aluno (medido em milissegundos);
- *Average Time Between Questions* (ATBQ): tempo médio que o aluno gastou entre cada questão (medido em milissegundos);
- *Decision Making Ratio* (DMR): rácio entre o número de respostas inseridas/alteradas/removidas e o número total de ações (medida percentual);
- *Correct Decision-Making Ratio* (CDMR): rácio entre o número de decisões consideradas corretas e o número de respostas inseridas/alteradas/removidas (medida percentual). Uma decisão é considerada correta quando um aluno insere/altera uma resposta para uma opção correta ou na remoção de uma resposta considerada incorreta;
- *Final Grade*: classificação final do aluno num exame após confirmação da finalização da tarefa (medida percentual).

Para a resolução do problema, deve começar por analisar e visualizar a distribuição das *features* do *dataset*, de modo a perceber relações entre o valor numérico PSS e cada uma das *features*. Com este conhecimento, técnicas de normalização e seleção de *features* poderão ser utilizadas para melhorar a *performance* do futuro modelo preditivo. De seguida, diferentes modelos preditivos deverão ser treinados e validados, como forma de comparação e seleção do modelo que apresente a melhor *performance*. Como extra, técnicas de *hyper-parameterization* poderão ser aplicadas para otimizar a forma como o modelo aprende.

Grupos com número **par** utilizarão o *dataset* com o conjunto de *features* associados à dinâmica do rato. Grupos com número **ímpar** utilizarão o *dataset* com o conjunto de *features* associados à tomada de decisão do aluno.

O objetivo final passa por desenvolver um modelo para prever o estado de stresse do estudante (PSS) de acordo com as *features* definidas.

Esta 2ª fase do trabalho prático compreende a entrega do código desenvolvido e do relatório, definindo todos os procedimentos aplicados e respetiva justificação da sua utilização, apoiando-se na demonstração dos resultados adquiridos.

Entrega

O código resultante da realização do trabalho prático e o respetivo relatório em formato digital .PDF (fase 2) deverão ser enviados por correio eletrónico para cesar.analide@di.uminho.pt e fgoncalves@algoritmi.uminho.pt, em ficheiros compactados (formato ZIP). Tanto o assunto da mensagem como o ficheiro deverão ser identificados na forma "[AEC: FYGXX]", em que [Y] designa o número da fase e [XX] designa o número do grupo de trabalho.

A sessão de apresentação do trabalho prático terá lugar no dia 3 de dezembro de 2018, em formato a anunciar oportunamente.

Comentado [FG1]: Verificar data de entrega!!!

Referências bibliográficas

Bowles, M. (2015). *Machine learning in Python: essential techniques for predictive analysis*. John Wiley & Sons.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc.

Gonçalves, F., Carneiro, D., Novais, P., & Pêgo, J. (2017, October). EUStress: A Human Behaviour Analysis System for Monitoring and Assessing Stress During Exams. In *International Symposium on Intelligent and Distributed Computing* (pp. 137-147). Springer, Cham.