 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Trabalho Prático para Avaliação em Período Letivo	Ano letivo 2022/2023	Data
	Curso Licenciatura em Engenharia Informática	Hora	
	Unidade Curricular Inteligência Artificial	Duração	

Registo de alterações:

- Versão inicial (1.0)

A. Destinatários

Este trabalho prático destina-se a todos os estudantes que pretendam ser avaliados durante o período letivo à Unidade Curricular de Inteligência Artificial da Licenciatura em Engenharia Informática.

O trabalho prático tem um peso de 50% na classificação final da UC e tem como requisito mínimo uma classificação de 8.0 valores

B. Objetivos

Este projeto funcionará como um elemento integrador dos conhecimentos adquiridos na UC de Inteligência Artificial, com um foco especial nas áreas de Computação Evolucionária e Machine Learning. Nomeadamente, serão trabalhadas competências fundamentais na área de IA, incluindo:

- A modelação do conhecimento existente no domínio de um problema e no seu espaço de solução, com vista à sua utilização computacional;
- A geração, otimização e avaliação de soluções válidas para um problema complexo, utilizando uma perspectiva evolucionária;
- A análise e comparação crítica de diferentes abordagens, com vista à seleção da mais adequada à resolução do problema;
- A melhoria iterativa e incremental de uma abordagem para a resolução de um problema com base em resultados passados;
- O treino, interpretação e avaliação de modelos de Machine Learning que resolvam um problema concreto.


C. Enunciado

O processo de encontrar um bom modelo de Machine Learning para um determinado problema é frequentemente complexo, em função da grande diversidade de algoritmos que existem hoje em dia, da quantidade de configurações que permitem, e da complexidade dos problemas a resolver em si.

Este processo pode ainda ser visto como um processo de otimização, em que o especialista em ML testa, iterativamente, diferentes algoritmos e/ou configurações, de forma a minimizar uma certa medida de erro, ou a maximizar uma certa métrica de performance, até obter um modelo considerado suficientemente bom para produção.

Nesse sentido, a utilização de um mecanismo de otimização para conduzir um processo de ML surge como uma evolução natural. De facto, este é uma das vertentes do Meta-Learning, uma das sub-áreas de Machine Learning. Existem diferentes abordagens de Meta-Learning mas uma delas pode ser descrita como “algoritmos que aprendem a aprender”.

Ou seja, enquanto um algoritmo de ML “aprende” um modelo com os dados, numa abordagem de Meta-Learning existe um outro processo que vai aprendendo a melhor forma de aprender, isto é, vai

 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Trabalho Prático para Avaliação em Período Letivo	Ano letivo 2022/2023	Data
	Curso Licenciatura em Engenharia Informática	Hora	
	Unidade Curricular Inteligência Artificial	Duração	

aprendendo que algoritmos e/ou que configurações vão dando melhores resultados, e vai fazendo a busca do modelo ideal evoluir nessa direção.

O principal objetivo deste trabalho prático é o de implementar um processo de Meta-Learning em que se usa um Algoritmo Genético para guiar o processo de busca pelo melhor modelo de Machine Learning. O problema específico a resolver é o da classificação de dígitos manuscritos, na área da Visão por Computador. Nesse sentido, será usado o dataset MNIST.

O dataset MNIST foi proposto em 1998 para a identificação de números escritos à mão. É um dos mais conhecidos datasets que representa um problema que é tipicamente fácil para os humanos de resolver, mas difícil para a “máquina”, sendo representado por uma base de dados de dígitos escritos à mão com um conjunto de treino composto por 60 000 exemplos (`train.csv`), e um conjunto de teste composto por 10 000 exemplos (`test.csv`).


Cada dígito é representado por uma matriz 28x28, totalizando 784 valores de pixels em escala de cinzento (features), que variam entre 0 (branco) e 255 (preto). O dataset contém ainda uma coluna (label) que contém um valor entre 0 e 9 que identifica o dígito representado. Trata-se, portanto, de um problema de ML supervisionado. A Figura 1 mostra alguns dos dígitos contidos no dataset. Como é possível notar, cada número pode ser escrito de formas muito diferentes, havendo ainda números que facilmente se confundem com outros, o que torna o problema complexo.



Fig. 1: Visualização de algumas instâncias do dataset MNIST.

Pretende-se, assim, que desenvolva um Algoritmo Genético que vá evoluindo diferentes modelos de Machine Learning para este problema, de forma a que vá encontrando soluções cada vez melhores para o problema da classificação de dígitos manuscritos.

Os resultados sucessivos que forem encontrados pelo Algoritmo Genético deverão ser submetidos numa competição a decorrer no Kaggle. Esta competição terá uma leaderboard atualizada em tempo real, que

 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Trabalho Prático para Avaliação em Período Letivo	Ano letivo 2022/2023	Data
	Curso Licenciatura em Engenharia Informática	Hora	
	Unidade Curricular Inteligência Artificial	Duração	

permitirá a cada equipa ter uma noção do seu desempenho relativo e absoluto. A posição nesta leaderboard terá ainda um peso no cálculo da nota final, tal como descrito mais abaixo. A competição está acessível a partir do link <https://www.kaggle.com/t/cb99330f6c1e45778b1f79131fdc422f>

Devem ser tidos em consideração os seguintes requisitos:

1. Devem ser considerados no mínimo 3 algoritmos diferentes;
2. Podem ser utilizadas as bibliotecas scikit-learn ou H2O. Fica ao critério de cada grupo usar apenas uma ou ambas;
3. O Algoritmo Genético deve ser tão configurável quanto possível;
4. Cada execução do Algoritmo Genético deve produzir evidências da sua performance ao longo do tempo (e.g. evolução da função de fitness ao longo do tempo, tempo de execução, etc.);
5. O trabalho deve ser desenvolvido utilizando a linguagem Python, e na forma de uma aplicação que possa ser executada com um único comando a partir da consola. Não obstante, durante o processo de desenvolvimento podem ser utilizadas outras alternativas tais como Jupyter Notebooks. Outras linguagens poderão ser utilizadas, mas nesse caso fica à responsabilidade de cada grupo utilizar bibliotecas de Machine Learning apropriadas,
6. A utilização de boas práticas de desenvolvimento software, incluindo a utilização de um repositório de código que permita aferir a evolução do trabalho ao longo do tempo e a contribuição de cada elemento do grupo, será valorizada;
7. A submissão de soluções na competição que decorrerá na plataforma Kaggle é de carácter obrigatório e servirá como um dos elementos de avaliação;
8. O processo evolucionário que permitiu chegar à melhor solução submetida na plataforma Kaggle deve ser descrito em detalhe no relatório, de forma a que seja possível determinar a autenticidade/veracidade dos resultados;

D. Recursos

Juntamente com este enunciado é fornecido um Jupyter Notebook que exemplifica o processo típico de Machine Learning utilizando as duas bibliotecas disponíveis para utilização neste trabalho prático (scikit-learn e H2O). No caso, utiliza-se um dataset com características de vinhos verdes (também disponibilizado em anexo a este enunciado), sendo o problema o de tentar “adivinhar” se um vinho é branco ou tinto pelas suas características.


O notebook fornecido exemplifica as diferentes fases do processo incluindo o carregamento dos dados, a sua divisão em conjuntos de treino, teste e validação, o treino do modelo, a avaliação da sua qualidade, e a sua utilização para fazer previsões.

E. Requisitos Mínimos/Critérios de Avaliação

Esta secção descreve os requisitos mínimos/obrigatórios para obtenção de aprovação à UC, bem como os critérios de avaliação. Nesse sentido, o relatório a desenvolver e a entregar juntamente com o projeto deve indicar claramente de que forma é que o grupo cumpriu ou não cada um destes requisitos.

É obrigatório o desenvolvimento de um Algoritmo Genético, com todos os operadores abordados em aula, sob a forma de uma aplicação, que evolua modelos de Machine Learning para o problema descrito acima.

É obrigatória a utilização de, pelo menos, 3 algoritmos diferentes na pesquisa dos modelos.

 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Trabalho Prático para Avaliação em Período Letivo	Ano letivo 2022/2023	Data
	Curso Licenciatura em Engenharia Informática	Hora	
	Unidade Curricular Inteligência Artificial	Duração	

É obrigatória a submissão de soluções na competição Kaggle.

A nota final da componente prática será obtida de acordo com a seguinte fórmula:

NFprática = 50% Desenvolvimento + 25% Qualidade do Relatório + 15% Qualidade da apresentação/defesa + 10% Quartil na leaderboard

Sugere-se que a equipa indique claramente no relatório os seguintes elementos, que servirão como critério de avaliação:

- Como foi implementada a função de fitness
- Como foi modelada uma solução
- Qual a lógica seguida para a implementação de cada elemento do Algoritmo Genético
- Quais os problemas encontrados e as estratégias implementadas para lidar com eles (e.g. lidar com o elevado tempo de treino dos modelos)
- Que estratégias de seleção/reprodução foram testadas e implementadas
- De que forma é que o Algoritmo Genético pode ser configurado
- Qual a configuração com a qual foi obtido o melhor desempenho, que deve incluir não apenas a qualidade do modelo mas ainda métricas de performance do próprio algoritmo genético (e.g. tempo de execução, número de gerações, tamanho da população)
- Utilização de gráficos e outros elementos que permitam perceber, entre outros, a evolução do fitness da população ao longo do tempo
- Outros elementos que a equipa considere importante

F. Relatório


Juntamente com o projeto, deverá ser entregue um relatório escrito que detalhe os aspetos mais relevantes do trabalho desenvolvido. O relatório será o elemento central na avaliação do trabalho. Entre outros elementos que o grupo considere relevantes, o relatório deve focar claramente e no mínimo os aspetos referidos na alínea E.

Cada seção do relatório deverá ser constituída por texto e excertos e código (quando aplicável e se imprescindível à explicação, sob a forma de imagens, legíveis) que detalhem aspetos essenciais do trabalho desenvolvido. Cada imagem (com a respetiva legenda) deverá ser devidamente enquadrada e mencionada no texto.

G. Realização

Este trabalho é realizado em equipas de 3 elementos, salvo situações particulares a validar previamente com o docente da UC.

Os alunos deverão constituir equipas e escolher um nome para a mesma. Este nome deverá ser utilizado como o nome da equipa na competição Kaggle. Deve ainda ser preenchido o ficheiro Excel com o nome “LEI.IA.22.23.Equipas”, disponível na secção “Files” do MSTeams de IA, colocando o nome da equipa em cada um dos seus membros.

 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Trabalho Prático para Avaliação em Período Letivo	Ano letivo 2022/2023	Data
	Curso Licenciatura em Engenharia Informática	Hora	
	Unidade Curricular Inteligência Artificial	Duração	

O trabalho desenvolvido deve ser entregue na página da UC no Moodle, até às 23:55 do dia 4 de Junho de 2023.

Apenas um elemento de cada equipa deverá submeter o trabalho em nome do grupo.

A apresentação e defesa do trabalho, de carácter obrigatório, decorrerá nas aulas do dia 6 de junho. Todos os elementos do grupo deverão estar presentes simultaneamente.

H. Elementos a entregar

Cada grupo deverá entregar um ficheiro .zip cujo nome segue a convenção IA_XXX.zip, em que XXX representa o nome da equipa. O .zip deverá ter o seguinte conteúdo:

- Relatório_IA_XXX.pdf – relatório em pdf, detalhando todos o trabalho desenvolvido e seguindo a estrutura sugerida
- Projeto_IA_XXX.zip – ficheiro .zip contendo o projeto desenvolvido em que se implementa o Algoritmo Genético que responde ao problema proposto neste enunciado
- Grupo_IA_XXX.txt – ficheiro de texto contendo a identificação (nome e número) dos elementos que constituem a equipa