**E-fólio B** | Folha de resolução para E-fólio

**UNIDADE CURRICULAR:** Raciocínio e Representação do Conhecimento

**CÓDIGO:** 21097

**DOCENTE:** José Coelho

**A preencher pelo estudante**

**NOME:** Gonçalo Caraça

**N.º DE ESTUDANTE:** 2000130

**CURSO:** Engenharia Informática

**DATA DE ENTREGA:** 29.05.2023

# *Índice*

# 1. INTRODUÇÃO

Neste trabalho, explorar-se-á o tema da aprendizagem, com foco nos diferentes tipos e técnicas utilizadas no campo da inteligência artificial e de aprendizagem máquinas. O objetivo é entender como os sistemas computacionais podem adquirir conhecimento, tomar decisões e melhorar seu desempenho ao longo do tempo.

## 1.1. Aprendizagem

No campo da Inteligência Artificial, a aprendizagem é uma área de estudo que procura capacitar sistemas computacionais a adquirirem conhecimento e melhorarem o seu desempenho ao longo do tempo. Através de algoritmos e modelos, esses sistemas são capazes de aprender a partir de dados, identificar padrões, tomar decisões e realizar tarefas de forma autônoma.

A aprendizagem pode ser dividida em diferentes tipos e abordagens, como aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Cada uma dessas abordagens possui características distintas e é aplicada em diferentes contextos, dependendo do problema a ser resolvido e dos dados disponíveis.

### 1.1.1. Aprendizagem Supervisionada

Neste tipo de aprendizagem, um modelo é treinado usando um conjunto de dados rotulados, em que cada exemplo de treino possui uma entrada e uma saída esperada. O objetivo é que o modelo aprenda a mapear as entradas para as saídas corretas. Durante o treino, o modelo ajusta os parâmetros para minimizar a diferença entre as saídas previstas e as saídas esperadas. Posteriormente, o modelo pode ser usado para fazer previsões ou classificar novos dados com base no conhecimento adquirido.

### 1.1.2. Aprendizagem Não Supervisionada

Neste tipo de aprendizagem, o modelo é exposto a um conjunto de dados não rotulados, sem informações prévias sobre as saídas esperadas. O objetivo é que o modelo encontre estruturas ou padrões ocultos nos dados, agrupando ou reduzindo as dimensões dos mesmos. Diferente da aprendizagem supervisionada, não há uma saída específica para o modelo aprender, sendo mais exploratório e descritivo. É frequentemente usado para descobrir informações úteis, como segmentação de clientes, deteção de anomalias, entre outros.

### 1.1.3. Aprendizagem Reforçada

Neste tipo de aprendizagem, o modelo aprende através d0 interação com um ambiente dinâmico. O modelo, chamado de agente, realiza ações num ambiente e recebe recompensas ou punições com base no seu desempenho. O objetivo do agente é aprender a tomar ações que maximizem as recompensas ao longo do tempo. A aprendizagem por reforço é frequentemente usada em jogos, robótica e tomada de decisões sequenciais, onde o agente aprende por tentativa e erro, através da exploração e explotação das ações disponíveis.

## *1.2. Classificação / Regressão*

A classificação e a regressão são duas tarefas comuns no campo da aprendizagem de máquinas e da análise de dados. São ambas técnicas de modelagem preditiva e tem como principal objetivo fazer previsões com base em dados históricos. No entanto, existem algumas diferenças entre elas.

A classificação é um problema de aprendizagem supervisionada em que o objetivo é atribuir uma categoria ou classe a uma observação com base nas suas características. É usado quando a variável de destino ou saída é categórica ou qualitativa. O modelo de classificação aprende com exemplos rotulados de treino e, em seguida, é capaz de prever a classe de novos exemplos não rotulados.

A saída de um modelo de classificação é uma classe ou categoria específica, como "sim" ou "não", "verdadeiro" ou "falso", ou uma das várias classes predefinidas. Os algoritmos comumente usados para classificação incluem árvores de decisão, algoritmos baseados em vizinhos mais próximos (k-NN), máquinas de vetores de suporte (SVM) e redes neuronais.

A regressão, por sua vez, também é um problema de aprendizagem supervisionada, mas é usado quando a variável de destino ou saída é contínua ou quantitativa. O objetivo da regressão é prever um valor numérico ou uma tendência com base nas características dos dados de entrada.

A saída de um modelo de regressão é um valor numérico contínuo. Alguns algoritmos populares de regressão incluem regressão linear, regressão logística, árvores de decisão de regressão, redes neuronais e regressão de floresta aleatória.

# 2. ANÁLISE AO DESAFIO PROPOSTO

O problema proposto consiste em explorar os dados do PORDATA, selecionar um objetivo de desenvolvimento sustentável e um indicador relacionado, e aplicar métodos de aprendizagem para prever ou analisar a relação entre o indicador selecionado e outros indicadores observáveis. A análise dos resultados permitirá entender melhor as interações e influências entre os indicadores.

## 2.1. Objetivo e indicadores selecionados

O objetivo escolhido foi o 17 - Parcerias para a implementação dos objetivos, com foco na previsão do Financiamento da União Europeia aos países em desenvolvimento. Os indicadores observáveis selecionados foram a taxa de risco de pobreza, as despesas em atividades de investigação e desenvolvimento (% do PIB) e a taxa de desemprego.

De entro todos os outros indicadores disponíveis para efetuar o treino, acredita-se que estes indicadores observáveis possam ter um impacto no objetivo da aprendizagem.

No entanto, é importante considerar que outros fatores externos, como políticas governamentais e condições econômicas, podem ter um impacto direto no financiamento, o que poderá tornar a previsão desafiadora.

## *2.2. Técnicas de aprendizagem*

Existem diversas técnicas de aprendizagem no campo da inteligência artificial e da aprendizagem de máquinas, sendo que cada uma delas possui características distintas, aplicações específicas e é mais adequada para diferentes tipos de problemas. No contexto do desafio proposto, optou-se por se utilizar as técnicas de Árvores de Decisão, k-Vizinhos Mais Próximos e Redes Neuronais.

### 2.2.1. Árvores de Decisão

As árvores de decisão são modelos populares e poderosos na aprendizagem de máquinas. Sua estrutura em forma de árvore facilita a interpretação das regras de decisão. Elas lidam com diferentes tipos de dados e podem ser aplicadas a problemas de classificação e regressão.

No problema proposto, as árvores de decisão poderão ser uteis, porque a visualização das árvores ajuda a interpretar os resultados, identificando padrões e relações entre os indicadores e o financiamento.

No entanto, as árvores de decisão são sensíveis a variações nos dados e podem se tornar complexas. Para melhorar seu desempenho, técnicas de poda ou combinação de várias árvores, como florestas aleatórias, podem ser utilizadas.

### 2.2.2. k-Vizinhos Mais Próximos

Os k-vizinhos mais próximos (KNN) é um algoritmo de aprendizagem de máquinas simples e intuitivo. Ele classifica ou faz previsões com base na proximidade dos exemplos de treino no espaço de características. O KNN seleciona os "k" vizinhos mais próximos de um exemplo de teste e atribui a ele a classe ou valor mais comum entre estes vizinhos.

Uma das principais vantagens do KNN é a sua simplicidade e facilidade de implementação. Além disso, ele pode lidar com dados complexos e não paramétricos. O KNN também é flexível, permitindo a classificação e regressão.

No entanto, o KNN também apresenta algumas desvantagens. Ele pode ser computacionalmente muito exigente, pois requer o cálculo da distância entre todos os exemplos de treino e teste. Além disso, o KNN é sensível à escolha do valor de "k" e à escala dos atributos, o que pode afetar a qualidade das previsões.

### 2.2.3. Redes Neuronais

As redes neuronais são um poderoso modelo de aprendizagem de máquinas inspirado no funcionamento do cérebro humano. Elas são compostas por camadas de neurônios interconectados, onde cada neurônio processa informações e transmite para os neurônios subsequentes. As redes neuronais têm a capacidade de aprender padrões complexos e extrair informações relevantes dos dados.

Uma das principais vantagens das redes neuronais é a sua capacidade de lidar com problemas não lineares e de alta complexidade. Elas podem capturar relações não triviais, além disso, elas podem-se adaptar aos dados, ajustando os pesos das conexões entre os neurônios durante o treino, o que as torna bastante flexíveis.

No entanto, as redes neuronais também têm algumas desvantagens. Elas exigem um grande volume de dados para um treino eficaz e podem ser computacionalmente muito exigentes. Além disso, a interpretação dos resultados

das redes neuronais pode ser desafiadora devido à sua natureza. Às vezes, é difícil entender como as decisões são tomadas dentro da rede.

# 3. RESULTADOS

De seguida, serão apresentados os resultados obtidos e a interpretação das técnicas aplicadas utilizando a linguagem R, uma ferramenta amplamente utilizada para análise de dados e implementação de modelos de aprendizagem de máquinas.

### 3.1.1. Tratamento de dados

O tratamento dos dados foi realizado por meio do pacote "readxl" para importar as tabelas do Excel que contem os dados necessários para a realização do projeto.

Posteriormente, foram aplicadas transformações específicas em cada tabela para formatar os dados. Isso incluiu a conversão das linhas em anos, a remoção de colunas desnecessárias, a seleção de dados a cada dois anos, a junção dos anos com os países, a exclusão de linhas com valores zero e a atribuição de nomes às colunas da tabela de objetivo.

Após o tratamento, as tabelas foram intersectadas com base nos nomes das colunas, resultando em uma tabela final contendo os dados dos indicadores observáveis e o indicador objetivo.

Em seguida, as tabelas foram unidas em uma única tabela final, que foi dividida em conjuntos de treino e teste usando a função "createDataPartition" do pacote "caret". A tabela de treino com 80% dos dados, e a tabela de teste com os restantes 20%.

Essa abordagem permitiu organizar os dados de forma adequada para o treino e teste dos modelos escolhidos, obtendo os seguintes valores para cada uma das tabelas:

| | |
|---|---|
| tabela_final | 127 obs. of 4 variables |
| tabela_teste | 24 obs. of 4 variables |
| tabela_treino | 103 obs. of 4 variables |

Exemplo de output das tabelas, neste caso específico da tabela final:

| | X1_Taxa_de_risco_de_pobreza_da_população_empregada | X2_Despesas_em_atividades_de_investigação_e_desenvolvimento_I_D_em_do_PIB | X3_Taxa_de_desemprego_de_longa_duração_por_sexo | X4_OBJETIVO_Financiamento_da_União_Europeia_aos_países_em_desenvolvimento |
|---|---|---|---|---|
| 2004:FR - França | 5.4 | 2.09 | 2.3 | 10128 |
| 2006:FR - França | 6.0 | 2.05 | 2.5 | 17783 |
| 2008:FR - França | 6.5 | 2.06 | 1.8 | 27632 |
| 2010:UE () - União Europeia (desde ) | 8.5 | 1.97 | 4.0 | 108002 |
| 2010:DE - Alemanha | 7.2 | 2.73 | 3.2 | 31407 |
| 2010:AT - Áustria | 7.5 | 2.73 | 1.5 | 4806 |
| 2010:BE - Bélgica | 4.5 | 2.06 | 3.7 | 5656 |
| 2010:DK - Dinamarca | 6.5 | 2.92 | 1.4 | 3616 |
| 2010:SK - Eslováquia | 5.7 | 0.61 | 10.7 | 56 |
| 2010:SI - Eslovénia | 5.3 | 2.05 | 3.2 | 44 |
| 2010:ES - Espanha | 10.9 | 1.36 | 7.3 | 7800 |
| 2010:FI - Finlândia | 3.7 | 3.71 | 2.5 | 3252 |
| 2010:FR - França | 6.5 | 2.18 | 2.4 | 26551 |
| 2010:GR - Grécia | 13.8 | 0.60 | 5.2 | 574 |
| 2010:HU - Hungria | 5.3 | 1.13 | 5.3 | 86 |
| 2010:IE - Irlanda | 5.5 | 1.59 | 6.9 | 2033 |
| 2010:IT - Itália | 9.5 | 1.22 | 4.3 | 7248 |
| 2010:LU - Luxemburgo | 10.6 | 1.42 | 1.3 | 310 |
| 2010:NL - Países Baixos | 5.1 | 1.70 | 1.1 | 9816 |
| 2010:PL - Polónia | 11.4 | 0.73 | 3.1 | 285 |
| 2010:PT - Portugal | 9.7 | 1.54 | 6.3 | 122 |
| 2010:CZ - República Checa | 3.7 | 1.33 | 3.0 | 172 |
| 2010:SE - Suécia | 7.7 | 3.17 | 2.5 | 3868 |
| 2012:UE () - União Europeia (desde ) | 8.9 | 2.08 | 4.9 | 98568 |
| 2012:DE - Alemanha | 7.8 | 2.88 | 2.4 | 27021 |
| 2012:AT - Áustria | 8.1 | 2.91 | 1.5 | 3734 |
| 2012:BE - Bélgica | 4.5 | 2.28 | 3.1 | 2104 |
| 2012:HR - Croácia | 6.0 | 0.74 | 10.2 | 17 |

### 3.1.2. Árvores de Decisão

Adaptando o script das Árvores de Decisão fornecido pelo professor para este caso específico, obteve-se os seguintes resultados:

```
                Number of trees: 10
No. of variables tried at each split: 1

        OOB estimate of  error rate: 100%
```

Pode-se observar que a taxa de erro da amostra (OOB) obtida é de 100%. Isso significa que o modelo não foi capaz de realizar previsões corretas para nenhum dos dados de teste.

Também ficou provado se tratar de uma regressão:

```
> teste$type                                  # confirmar que foi uma classificação e não uma regressão
[1] "regression"
```

Sendo assim pode-se concluir que devido à natureza dos dados utilizados, que são influenciados por fatores externos relacionados à política, economia europeia e interesses econômicos, é importante considerar que esses fatores não podem ser controlados no treino do modelo. Uma vez que se tem acesso apenas aos dados disponibilizados pelo PORDATA e tratando-se de um

financiamento aos países em desenvolvimento, esses fatores externos exercem uma influência significativa no processo de treino.

Dado este contexto, uma abordagem que poderia melhorar o modelo seria realizar a classificação dos dados. Ao classificar os dados, podemos identificar padrões e relações que podem não ser imediatamente óbvios. Essa abordagem pode ajudar a capturar de forma mais eficiente as influências dos fatores externos e a melhorar a capacidade do modelo.

Foi então separado a coluna com o indicador objetivo em 4 classes repartidas em quantias iguais, obtendo o seguinte resultado:

| | X1_Taxa_de_risco_de_pobreza_da_população_empregada | X2_Despesas_em_atividades_de_investigação_e_desenvolvimento_I_D_em_do_PIB | X3_Taxa_de_desemprego_de_longa_duração_por_sexo | categories |
|---|---|---|---|---|
| 2010:DK - Dinamarca | 6.5 | 2.92 | 1.4 | Class 3 |
| 2010:SI - Eslovénia | 5.3 | 2.05 | 3.2 | Class 1 |
| 2010:LU - Luxemburgo | 10.6 | 1.42 | 1.3 | Class 2 |
| 2010:PL - Polónia | 11.4 | 0.73 | 3.1 | Class 2 |
| 2012:DE - Alemanha | 7.8 | 2.88 | 2.4 | Class 4 |
| 2012:SK - Eslováquia | 6.2 | 0.79 | 10.9 | Class 1 |
| 2012:HU - Hungria | 5.7 | 1.25 | 4.8 | Class 1 |
| 2012:PL - Polónia | 10.4 | 0.89 | 4.2 | Class 2 |
| 2012:CZ - República Checa | 4.5 | 1.77 | 3.0 | Class 1 |
| 2014:BE - Bélgica | 4.8 | 2.37 | 3.9 | Class 3 |
| 2014:BG - Bulgária | 9.2 | 0.79 | 7.4 | Class 1 |
| 2014:SI - Eslovénia | 6.4 | 2.37 | 5.3 | Class 1 |
| 2014:FR - França | 8.0 | 2.23 | 3.1 | Class 4 |
| 2014:MT - Malta | 5.5 | 0.69 | 2.9 | Class 1 |
| 2014:PT - Portugal | 10.7 | 1.29 | 8.7 | Class 2 |

Executando novamente o script agora para os dados classificados:

```
> teste$type                                  # confirmar que foi uma classificação e não uma regressão
[1] "classification"
```

Obteve-se os seguintes resultados:

```
OOB estimate of  error rate: 39.81%
```

Ao comparar o resultado atual com o resultado anterior, pode-se observar que houve uma melhoria significativa na taxa de erro da amostra (OOB).

Pode-se assim concluir que, após aplicar a abordagem de classificação aos dados, a taxa de erro OOB foi reduzida para 39.81%. Embora ainda haja uma taxa significativa de erro, essa melhoria indica que o modelo agora é capaz de fazer previsões mais precisas em comparação com a abordagem anterior, embora ainda longe de valores considerados aceitáveis.

De seguida foi calculada a precisão do resultado do treino para determinar a percentagem de acerto do modelo:

```
# Precisao do treino
precisao_resultado_treino_arvores<-sum((tabela_teste$categories==resultado_treino_arvores$`predict(teste, newdata = tabela_teste[, 1:3])`))/nrow(tabela_teste)
```

E obteve-se uma precisão de aproximadamente 0.708, o que indica que o modelo atingiu uma taxa de acerto de aproximadamente 70,8% das previsões com base nos dados de treino, isso indica que embora o resultado seja aceitável, ainda existe espaço para melhorias.

| | |
|---|---|
| precisao_resultado_treino_arvores | 0.708333333333333 |

### 3.1.3. k-Vizinhos Mais Próximos

Adaptando o script fornecido para a situação específica e tendo em conta que que agora se está a trabalhar com classificações, obteve-se o seguinte resultado:

```
Class 1 Class 2 Class 3 Class 4
     29      23      22      29
```

Estes resultados indicam a contagem de vizinhos mais próximos que foram atribuídos a cada classe com base nas características dos dados de treino. Nesse caso, temos 29 vizinhos mais próximos atribuídos à Class 1, 23 à Class 2, 22 à Class 3 e 29 à Class 4.

Precisão do treino obtido:

| | |
|---|---|
| precisao_resultado_treino_k_vizinhos | 0.666666666666667 |

Uma precisão de 66.66% indica que o modelo foi capaz de prever corretamente aproximadamente dois terços dos testes com base no treino efetuado, mas ainda há uma margem significativa de erro, e seria indicado utilizar outras técnicas para tentar melhorar os resultados, como por exemplo, considerar diferentes valores de k ou ajustar os parâmetros do modelo.

### 3.1.4. Redes Neuronais

Adaptando o script fornecido para o problema específico e tendo em conta, obteve-se os seguintes resultados:

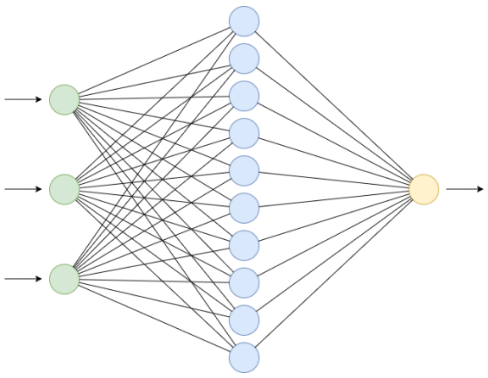
```
iter 400 value 6.702373
iter 410 value 6.698560
iter 420 value 6.697005
iter 430 value 6.696004
iter 440 value 6.695913
iter 450 value 6.695894
iter 460 value 6.695886
final  value 6.695886
converged
```

O output indica que o treino da rede neuronal foi executado com sucesso, atingindo o valor final de 6.696 com 460 iterações, o que indica uma redução significativa do erro em relação ao valor inicial, que era 1189.178.

Obteve-se uma arquitetura de rede:

```
3-10-1 network with 54 weights
```

Que indica que a arquitetura da rede neuronal contem 3 neurônios na camada de entrada, que são os 3 indicadores observáveis escolhidos, 10 neurônios na camada oculta e 1 neurônio na camada de saída.



Quanto à precisão do treino, o valor obtido foi de 0.542, o que corresponde a 54.2%. Isso significa que o modelo acertou pouco mais de metade das previsões do conjunto de treino.

| | |
|---|---|
| precisao_redes_neuronais | 0.541666666666667 |

Tendo em conta os resultados obtidos pode-se concluir que os resultados ficam bastante aquém dos aceitáveis, e que deveriam ser adotadas outras técnicas para tentar melhorar os resultados, como por exemplo, avaliar e ajustar a arquitetura da rede, mais dados ou até mesmo considerar outras técnicas.

### 3.1.5. Conclusões finais

Considerando os resultados obtidos nas técnicas utilizadas, pode-se concluir que todas elas se encontram ainda longe de um objetivo considerado aceitável, é por isso fundamental avaliar diferentes aspetos do problema e das técnicas utilizadas. Pode ser necessário explorar outras abordagens e/ou ajustar parâmetros para melhorar a precisão dos modelos. Além disso, pode ser útil realizar uma análise mais aprofundada dos dados, considerar a inclusão de novas variáveis relevantes e explorar outras técnicas de aprendizagem de máquinas.

É fundamental lembrar que o processo de modelagem e previsão é iterativo, e é comum que seja necessário experimentar diferentes abordagens para encontrar o modelo mais adequado e alcançar uma precisão satisfatória. Portanto, é recomendável continuar-se com a exploração dos modelos e das técnicas utilizadas até que se obtenha uma precisão que atenda aos requisitos e objetivos do problema em questão.

# *4. BIBLIOGRAFIA*

Artificial Intelligence: A Modern Approach, Stuart Russell, Peter Norvig, Prentice-Hall.

Materias disponibilizados na UC

**FIM**