Aplicação de Técnicas de Aprendizagem Automática para Analisar o Desempenho de Ciclistas

Ana Albergaria
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1201518@isep.ipp.pt

Rita Ariana Sobral
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1201386@isep.ipp.pt

Vasco Azevedo
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1202016@isep.ipp.pt

***Abstract*—Este artigo foi concebido no âmbito da unidade curricular de Análise de Dados em Informática, e teve como objetivo realizar um estudo do desempenho de diferentes modelos de previsão e classificação de aprendizagem automática, relativos a dados referentes à preparação de ciclistas para uma nova temporada de competições, após o treino de inverno.**

***Index Terms*—Análise de Dados, Inteligência Artificial, Machine Learning, Regreesão, Classificação**

I. Introdução

O artigo científico, "Aplicação de Técnicas de Aprendizagem Automática para Analisar o Desempenho de Ciclistas", foi realizado no âmbito da Unidade Curricular (UC) de Análise de Dados em Informática (ANADI), integrante no 6º semestre da Licenciatura em Engenharia Informática (LEI) do Instituto Superior de Engenharia do Porto (ISEP) e segue as diferentes fases de desenvolvimento de modelos de aprendizagem automática.

Os dados presentes no ficheiro fornecido pelos docentes da UC (ciclismo.csv) são relativos à preparação de ciclistas para uma nova temporada de competições, após o treino de inverno e foram analisados e processados com recurso à ferramenta RStudio e com a linguagem R.

Objetivos

Este artigo teve como principal objetivo a aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação usando os testes estatísticos mais adequados, de maneira a validar as possíveis relações entre os dados recolhidos e as medições realizadas para várias métricas que mostram o desempenho do atleta.

De maneira a responder de maneira positiva ao problema anteriormente enunciado, foram identificados os seguintes objetivos específicos:

- Definir a metodologia de trabalho
- Análise e Discussão dos Resultados com recurso ao R
- Escrita de artigo científico

Metodologia de Trabalho

A equipa decidiu adotar a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) [1] como base para a execução deste trabalho. Para garantir uma abordagem estruturada e eficaz, seguimos as seis fases principais dessa metodologia durante todo o processo.

II. Estado de Arte

Inteligência Artificial

A Inteligéncia Artificial centra-se no desenvolvimento de sistemas e computadores com vista a realizar tarefas que requerem inteligência humana [2].

É uma tecnologia multidisciplinar onde existem diversas abordagens e técnicas tais como Machine Learning, Redes Neuronais, Processamento de Linguagem Natural, Visão Computacional, entre outros [2].

Machine Learning

Machine Learning é uma subárea da Inteligência Artificial com vista no desenvolvimento de algoritmos e modelos estatísticos que permitem que os computadores realizem a análise e interpretação automática de dados, a identificação de padrões e previsões ou tomada de ações com base nesses padrões [3].

Existem três métodos principais de Machine Learning:

- **Aprendizagem Supervisionada** - a cada saída, isto é, parâmetro a avaliar, é atribuído um rótulo (valor numérico ou classe). O algoritmo é treinado com base em valores de entradas e saídas conhecidas de forma a que, posteriormente, seja capaz de prever o rótulo de saída com base na entrada recebida [4].
  Se o rótulo de saída for um valor real, trata-se de um algoritmo de regressão [4].
  Se a saída poder assumir apenas um conjunto de rótulos pré-definidos, estamos perante um algoritmo de classificação [4].

- **Aprendizagem Não Supervisionada** - não é atribuído um rótulo para os dados de saída. O algoritmo é responsável por identificar padrões a partir de um conjunto grande de dados, de forma a identificar grupos de itens semelhantes [4].
- **Aprendizagem por Reforço** - existem dois componentes principais: o agente e o ambiente. O agente aprende por tentativa e erro, através da obtenção de feedback sob a forma de recompensa (ou penalidade) após realizar determinada ação, com vista a maximizar esse sinal de recompensa [4].

Regressão Linear

A Regressão Linear é um modelo de estatística cujo objetivo centra-se em explicar a relação entre uma variável Y, variável dependente a prever designada por resposta, e uma ou mais variáveis independentes ditas preditoras X1,...,Xp, em que p >= 1. [5]

Regressão Linear Simples

A Regressão linear é simples quando existe uma relação linear entre a variável dependente Y e uma variável independente preditora X. [5]

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Regressão Linear Múltipla

A Regressão linear é múltipla quando existe uma relação linear entre a variável dependente e um conjunto de variáveis independentes preditoras X1,...,Xp (p >= 1). [5]

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon$$

Árvore de Regressão

A árvore de regressão é um modelo que resolve problemas de regressão, isto é, a variável que se deseja prever é contínua. Como o nome indica, possui uma estrutura em árvore de nós e raamificações. Cada nó interno representa um atributo da instância, e cada raamificação representa um resultado possível. As folhas da árvore contêm os valores previstos de regressão [6].

Árvore de Decisão

A árvore de decisão segue, também, uma estrutura de árvore, porém é usada para problemas de classificação ou decisão, ou seja, a variável alvo é uma variável categórica. Cada nó interno representa uma condição que divide o conjunto de dados, e cada ramificação representa uma escolha entre várias categorias possíveis [6].

Rede Neuronal

A rede neuronal é um algoritmo baseado nas redes de neurónios que compõem o sistema nervoso central humano [7].

É agrupada, normalmente, em três (ou mais camadas): camada de entrada, camadas escondidas (hidden layers), e camada de saída [7].

Cada nó conecta-se a outro e possui um peso e limite associados. O nó é ativado e envia dados para a próxima camada da rede, se a sua saída for superior ao valor limite definido. Caso contrário, nenhum dado é enviado para a próxima camada da rede [7].

A rede neuronal é treinada usando o algoritmo de backpropagation, cujo objetivo é ajustar os pesos das ligações entre os nós de forma a reduzir o erro entre as saídas obtidas e as saídas desejadas durante o treino [8].

K-vizinhos-mais-próximos.

O algoritmo K-vizinhos-mais-próximos tem como principal objetivo prever a classe de um determinado elemento tendo como base a classe dos vizinhos k mais próximos desse elemento [4].

O k deve ser escolhido com precaução, uma vez que escolher um k demasiado pequeno ou demasiado alto traz implicações. A escolha de um k demasiado baixo pode provocar a presença de ruído - desvio significativo - nos dados de treino. Por outro lado, a escolha de um k demasiado alto pode causar a suavização excessiva das fronteiras de decisão [4].

É um algoritmo *lazy learning*, pelo que não requer treino do modelo antes de efetuar previsões. Ou seja, os dados de treino são armazenados e no memento de realização de uma previsão, estes são recuperados para serem utilizados. Por essa razão, acarreta algumas implicações [9]:

- São mais lentos que os restantes algoritmos, como os algoritmos *eager learning*
- São também mais sensíveis a ruídos e a *outliers*, o que pode causar uma menor precisão dos dados
- Caso o conjunto de dados de treino seja vasto, requer um grande espaço em memória, visto que estes são armazenados para serem utilizados na fase de teste
- Menor capacidade de generalização e identificação de padrões nos dados

Avaliação de Desempenho

Matriz de Confusão

A matriz de confusão é uma tabela com a finalidade de avaliar o desempenho de um modelo de classificação sobre um conjunto de dados, onde são apresentadas as classes reais e previstas.

	Negativo Previsto	Positivo Previsto
Negativo Atual	TN	FP
Negativo Previsto	FN	TP

As métricas de desempenho que avalia são as seguintes:

- **Accuracy** - capacidade de um classificador binário de classificar acertadamente tanto os positivos como negativos
- **Precision** - proporção dos positivos previstos que são verdadeiramente positivos
- **Sensitivity** - capacidade de um classificador binário detetar verdadeiros positivos (proporção de valores realmente positivos classificados corretamente)

- **Specificity** - capacidade de um classificador binário de detetar verdadeiros negativos.
- **F1** - valor entre 0 e 1 que corresponde à média harmónica da precisão e da sensibilidade

Hold-out

A técnica designada por holdout consiste na divisão dos dados em treino e teste, dada alguma proporção pré-definida. O algoritmo aprende com o subconjunto de treino e o restante dos dados é usado para predição.

K-Fold Cross Validation

O método K-Fold Cross-Validation avalia a performance do modelo em diferentes subdivisões dos dados de treino [10].

É um método robusto para estimar a taxa de acerto de um modelo [10].

III. Realização do Projeto

Análise e Tratamento dos Dados

O conjunto de dados fornecido, ciclismo.csv, contém informações sobre os treinos realizados no período de pré-temporada por ciclistas profissionais.

De maneira a obter uma análise descritiva do *dataset*, recorremos a algumas funções do R tais como "dim", "summary" e "str". Estas funções fornecem várias informações úteis tais como as dimensões do dataset, tendo, portanto, 11 colunas, cada uma destas com 1000 linhas, como é visível na Fig. 1

```
'data.frame':   1000 obs. of  11 variables:
 $ ID                : chr  "0" "1" "2" "3" ...
 $ gender            : chr  "female" "male" "male" "female" ...
 $ Team              : chr  "group D" "group D" "group E" "group E" ...
 $ Background        : chr  "Sprinter" "None" "Cobblestones" "Mountain" ...
 $ ProLevel          : chr  "Continental" "World Tour" "World Tour" "Continental" ...
 $ WinterTrainingCamp: chr  "completed" "none" "none" "none" ...
 $ altitude_results  : chr  "44" "75" "78" "56" ...
 $ vo2_results       : chr  "56" "68" "60" "67" ...
 $ hr_results        : chr  "55" "69" "60" "68" ...
 $ dob               : chr  "1988-01-09" "1998-07-22" "1995-03-19" "2003-04-27" ...
 $ Continent         : chr  "North America" "Africa" "Africa" "Australia" ...
```

Fig. 1. Informações sobre o conjunto de dados

Ao analisar a Fig. 1 reparámos no atributo dob que representa a data de nascimento dos ciclistas, mas essa informação, devido à maneira que estava representada, poderia não se traduzir em informação relevante para os modelos e por este motivo criamos um novo atributo Age que é derivado através do atributo dob e que representa a idade atual do ciclista.

Para um melhor entendimento sobre os atributos do dataset foi realizada uma análise dos dados que entendemos que seriam mais significativos.

Fizemos uma análise dos atributos categóricos, para averiguar se existiam atributos binários e assim os representar como valores de 0 ou 1, uma vez que esta alteração pode ter impacto no desempenho dos modelos. Esta análise identificou que os atributos gender, ProLevel e WinterTrainingCamp poderiam sofrer esta alteração.

De seguida, realizamos uma análise aos atributos numéricos do dataset. Começamos por criar gráficos "box-plot" entre as diferentes variáveis para identificar como estariam relacionadas. Verificámos que a idade não apresenta nenhuma relação visível com as restantes variáveis numéricas. Em oposição as variáveis altitude_results, vo2_results e hr_results apresentam uma elevada relação, visto que à medida que uma aumenta as outras também acompanham esse aumento.

Depois de analisados os dados, é importante realizar um pré-processamentos dos dados, tal como a identificação de outliers e a selação de atributos.

Inicialmente procedeu-se à remoção de todas as linhas que possuíam valores NA's, caso existissem.

De seguida, como é visível na Fig. 2, gerou-se um gráfico box-plot de maneira a determinar se existiam outliers nos dados das variáveis altirude_results, hr_results e vo2_results.

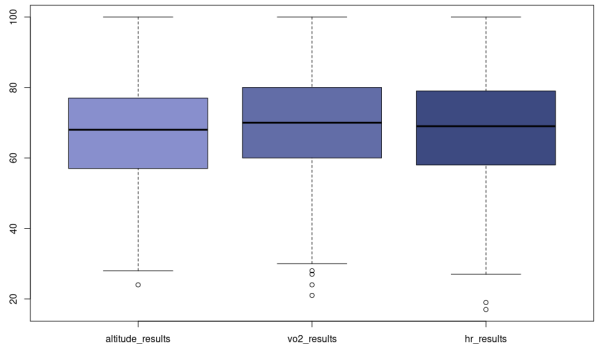

Fig. 2. Gráfico box-plot para altirude_results, hr_results e vo2_results

Visto que nenhum outlier parecia um erro de medição, logo eram valores possíveis de acontecer, as linhas que continham estes valores não foram retiradas.

As variáveis que não eram necessárias para o caso de estudo devido às suas características eram o Id e a Data de Nascimento, sendo, portanto, removidas.

Por fim foi realizada a codificação das variáveis categóricas que não eram binárias.

Matriz de Correlação

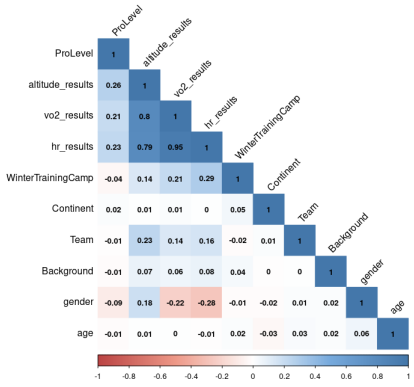

Fig. 3. Diagrama de correlação entre as variáveis

Através da análise do diagrama presente na Fig. 3 conseguimos concluir que existe uma correlação positiva entre vo2_results, altitude_results e hr_results, o que significa que se uma aumentar, as outras também aumentam, indo de encontro à análise realizada anteriormente,

De realçar que a maioria das variáveis estão compreendidas entre 0.25 e -0.25, logo a grande parte das variáveis não apresenta correlação.

Determinar a Variável altitude_results

Modelo de Regressão Linear Simples

Um parâmetro crucial ao criar modelos de regressão linear e outros algoritmos de Machine Learning (ML) é o seed, pois permite controlar a aleatoriedade presente durante o processo de treino e validação do modelo. Isso permite alcançar a reprodutibilidade do modelo e comparações mais precisas com outros modelos. Como resultado, começamos por definir o seed no início do processo.

Para garantir que o modelo fosse validado corretamente, adotamos a abordagem Holdout que consiste em dividir o dataset em 70% para treino, os dados utilizados para a criação do modelo, e os restantes 30% para teste. Esta divisão não deve enviezar os dados e isso foi verificado que não acontecia, com recurso à função ”summary” do R.

A variável altitude_results, variável dependente, foi determinada através da sua relação com a variável hr_results, variável independente, e o modelo de regressão linear simples resultante foi o seguinte:

$$altitude\_results = 14,4220 + 0,7636 hr\_results$$

Para fazer uma análise visual do modelo recorreu-se à visualização da reta correspondente ao modelo e ao respetivo diagrama de dispersão visível na Fig. 4, e como era espectável os valores do resultado do treino de altitude aumentam consoante o aumento do teste de frequência cardíaca.

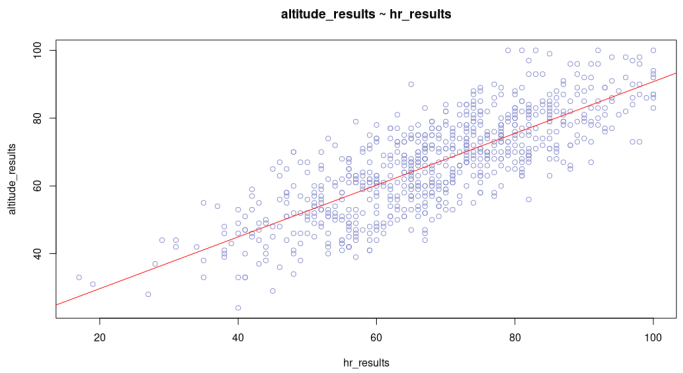

Fig. 4. Reta de Regressão Linear do Modelo Criado

Efetuou-se também o “summary” para o modelo de maneira a analisar o p-value obtido para a variável preditora, “hr_results”, sendo este de 2e-16, logo podemos considerá-la uma boa variável preditora. Relativamente ao R2, este assume o valor de 0.63, logo o modelo consegue explicar 63% da variabilidade total.

Com os restantes 30% dos dados, foi realizado um teste às capacidades preditivas do modelo de regressão linear. O erro absoluto médio (MAE) foi de 7,1413 e a raiz quadrada do erro quadrático médio (RMSE) foi de 8,5380. Tendo em conta que os valores de erro abrangem uma margem menor do que a amplitude interquartil de 20, podemos assumir que eles não são muito significativos.

Depois de analisado o modelo de regressão linear simples tentamos perceber se seria possível obter melhores resultados utilizando modelos mais complexos, tais como o modelo de regressão linear múltipla, a árvore de regressão e a rede neuronal, tendo todos os modelos utililizado a abordagem de Hold-out.

Modelo de Regressão Linear Múltipla

Depois de elaborado o modelo de regressão linear múltipla que relacionava o valor do treino de altitude com todos os outros atributos do dataset é necessário interpretar o sumário do modelo, nomeadamente o p-value que é menor que 2.2e-16 e o valor de R2 que é 0,84, o que significa que, pelo menos, uma das variáveis preditoras é significativamente relacionada com a variável de estudo e que o modelo consegue explicar 84% da variabilidade total.

Foi também obtido um sumário dos resultados de cada variável preditora, Fig. 5, para averiguar a existência de variáveis que não são estaticamente significativas.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.095419   1.593518  -3.198  0.00145 **
gender            12.325485   0.447517  27.542  < 2e-16 ***
Team               1.171149   0.192677   6.078 2.00e-09 ***
Background         0.083214   0.127147   0.654  0.51303
ProLevel           2.788189   0.463676   6.013 2.94e-09 ***
WinterTrainingCamp -2.658074   0.492769  -5.394 9.45e-08 ***
vo2_results        0.318410   0.050219   6.340 4.12e-10 ***
hr_results         0.581636   0.050181  11.591  < 2e-16 ***
Continent          0.132842   0.127081   1.045  0.29623
age               -0.004538   0.036895  -0.123  0.90215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 5. Sumário do resultado das variáveis preditoras

Através da análise da Fig. 5 conseguimos determinar que as oscilações de valores da variável altitude_results estão significativamente relacionadas com a variável hr_results e com a variável gender, existindo ainda mais algumas variáveis preditoras consideradas extremamente relevantes. Por outro lado, concluímos que as variáveis Background e Age não são estaticamente significativas.

Depois da análise efetuada foi elaborado outro modelo de regressão linear múltipla que relaciona o valor do treino de altitude com os atributos que são estaticamente significativos.

Este modelo foi o que conseguiu obter melhores resultados a determinar a variável “Altitude_results”, sendo o erro absoluto

médio (MAE) de 4,4386 e a raiz quadrada do erro quadrático médio (RMSE) de 5,4855.

Determinar a Variável vo2_results

Todos os modelos de regressão criados para determinar a variável vo2_results utilizam a abordagem Hold-out, onde é criado um conjunto de treino com 70% dos dados, e um de teste com 30%.

Modelo de Regressão Linear Múltipla

Depois de elaborado o modelo de regressão linear múltipla que relacionava a variável vo2_results com todos os outros atributos do dataset é necessário interpretar o sumário do modelo, nomeadamente o p-value que é menor que 2.2e-16 e o valor de R2 que é 0,91, o que significa que, pelo menos, uma das variáveis preditoras é significativamente relacionada com a variável de estudo e que o modelo consegue explicar 91% da variabilidade total.

Através da análise da Fig. 5 conseguimos determinar que as oscilações de valores da variável vo2_results estão significativamente relacionadas com a variável hr_results, sendo esta a variável preditora mais influente.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.70825    1.15880   4.926 1.05e-06 ***
gender            -0.94092    0.47395  -1.985 0.047510 *
Team              -0.49233    0.14401  -3.419 0.000666 ***
Background        -0.12761    0.09328  -1.368 0.171733
ProLevel          -0.86379    0.34773  -2.484 0.013223 *
WinterTrainingCamp -1.93701    0.36200  -5.351 1.19e-07 ***
altitude_results   0.17173    0.02708   6.340 4.12e-10 ***
hr_results         0.79216    0.02679  29.568  < 2e-16 ***
Continent          0.16670    0.09319   1.789 0.074073 .
age                0.03740    0.02706   1.382 0.167404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 6. Sumário do resultado das variáveis preditoras

Árvore de Regressão

A árvore de regressão foi elaborada com recurso à função rpart do R usando “anova”, tendo obtido a árvore de regressão presente na Fig. 7.

Através da análise da Fig. 7, conseguimos conluir que existe uma correlação direta significativa com a variável hr_results, sendo esta a mesma conclusão obtida depois de analisar o modelo de regressão linear múltipla.

Rede Neuronal

Com recurso à função neuralnet, elaborou-se o modelo da rede neuronal com o vo2_results para a totalidade dos dados de treino normalizados, e escolheu-se o número de nós como 1, sendo que foi feita uma avaliação de qual seria a melhor arquitetura, e esta foi considerada a mais eficiente, Fig. 8

	MAE	RMSE
Regressão linear múltipla	3,3396	4,1598
Árvore de regressão	4,0569	5,1440
Rede neuronal	3,4739	4,231

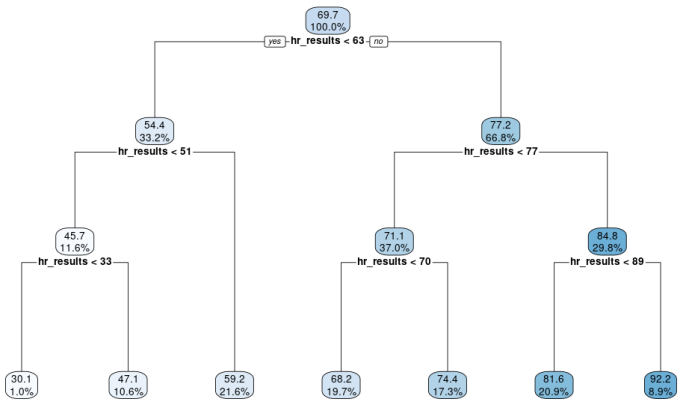

Fig. 7. Árvore de Regressão para o vo2_results

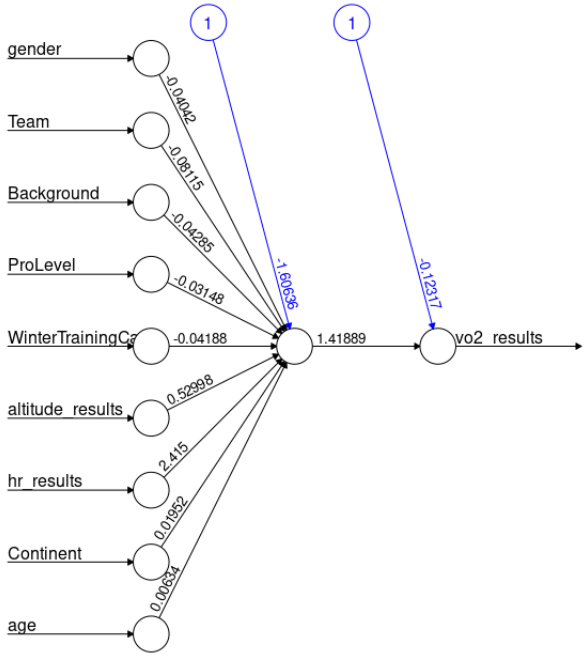

Fig. 8. Rede neuronal vo2_results

Tendo em conta os resultados dos erros dos modelos anteriores, a regressão múltipla e a rede neuronal foram considerados os modelos de regressão mais precisos.

Recorreu-se a testes de hipótese para averiguar se os resultados obtidos eram estatisticamente significativos. Para tal, foram comparados os MAE e RMSE dos dois melhores modelos e, como o p-value foi superior ao nível de significância, conclui-se que os resultados obtidos não são estatisticamente significativos. Contudo, o melhor modelo é a Regressão Linear Múltipla, uma vez que apresenta valores de MAE e RMSE menores.

Atributo Pro Level

Nesta fase do caso prático, foi necessário elaborar 3 métodos de classificação, e estudar a capacidade preditiva relativamente ao atriburo Pro Level.

Assim como nos modelos anteriores, foi utilizada uma seed de maneira a garantir a reprodutibilidade dos resultados, juntamente com o método holdout para dividir os dados em conjuntos de treino e teste.

Árvore de Decisão

Através da análise da árvore de decisão gerada, Fig. 9, conseguimos concluir que quando o resultado do treino de altitude é maior que 56, existe uma grande relação relativamente ao Pro Level ser Continental.

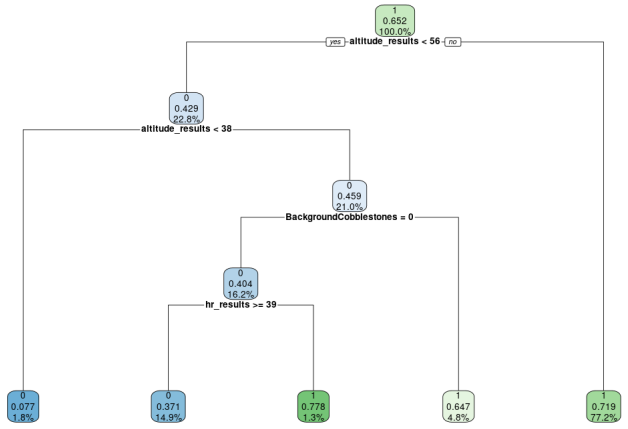

Fig. 9. Árvore de Decisão Pro Level

Já com a árvore gerada foi feita a previsão dos dados e obteve-se uma *accuracy* de aproximadamente **68,47%**.

Rede Neuronal

O modelo para a rede neuronal foi elaborado com uma arquitetura de 3 por 2, Fig. 10.

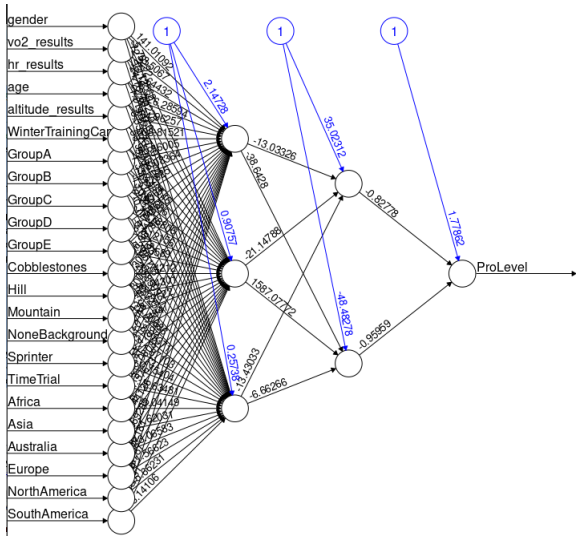

Fig. 10. Rede Neuronal Pro Level

Já com a rede neuronal gerada obteve-se uma *accuracy* de aproximadamente **67,46%**, tendo sido realizada uma aproximação dos valores previstos para o atributo Pro Level para os números inteiros mais próximos (i.e. 0 ou 1).

K-vizinhos-mais-próximos

No modelo k-vizinhos mais próximos, realizou-se um estudo sobre qual era o valor de K em que se obtinha a maior *accuracy* possível, tendo chegado à conclusão que o melhor valor era o 15 com uma *accuracy* de aproximadamente **96.95%**.

Depois de analisados os 3 modelos conseguimos concluir que os dois modelos com melhor capacidade preditiva são a Árvore de Decisão e o K-vizinhos-mais-próximos, tendo sindo utilizado o método k-fold cross validation para obter uma média e um desvio padrão da *accuracy* da previsão do atributo Pro Level, para os dois modelos enunciados anteriormente. Os resultados encontram-se na tabela a seguir.

	Média	Desvio Padrão
Árvore de Decisão	0.6844	0.0295
KNN	0.9722	0.0194

Através da análise dos valores obtidos para a média e desvio padrãos dos dois modelos conseguimos concluir que o modelo K-vizinhos-mais-próximos apresenta uma maior *accuracy* e mais consistência a prever o atributo Pro Level do que a Àrvore de Decisão.

Recorreu-se a testes de hipótese para a avaliação das diferenças entre todos os valores de *accuracy* calculados anteriormente, concluindo que existem diferenças significativas no desempenho dos dois melhores modelos obtidos anteriormente, sendo o modelo K-vizinhos-mais-próximos o melhor.

Para não tornar exaustiva a comparação dos 3 modelos, resumiu-se os valores obtidos para cada critério (Accuracy, Sensitivity, Specificity e F1) na tabela seguinte:

	Árvore de Decisão	Rede Neuronal	KNN
Accuracy	68,47%	67,46%	96.95%
Sensitivity	0.27	0.41	0.91
Specificity	0.89	0.79	1
F1	0.37	0.45	0.95

Os valores obtidos reforçam a conclusão obtida anteriormente, pois na totalidade dos critérios, existe vantagem para o modelo de K-vizinhos-mais-próximos.

Atributo Winter Training Camp

Foram elaborados 2 métodos de classificação com a finalidade de estudar a capacidade preditiva relativamente ao atributo Winter Training Camp.

Árvore de Decisão

Através da análise da árvore de decisão gerada, Fig. 11, conseguimos concluir que quando o resultado do teste de frequência cardíaca é menor que 78, existe uma grande relação relativamente ao Winter Training Camp não estar completo.

Já com a árvore gerada foi feita a previsão dos dados e obteve-se uma *accuracy* de aproximadamente **67.8%**.

Rede Neuronal

O modelo para a rede neuronal foi elaborado com uma arquitetura com 1 nó interno, Fig. 13.

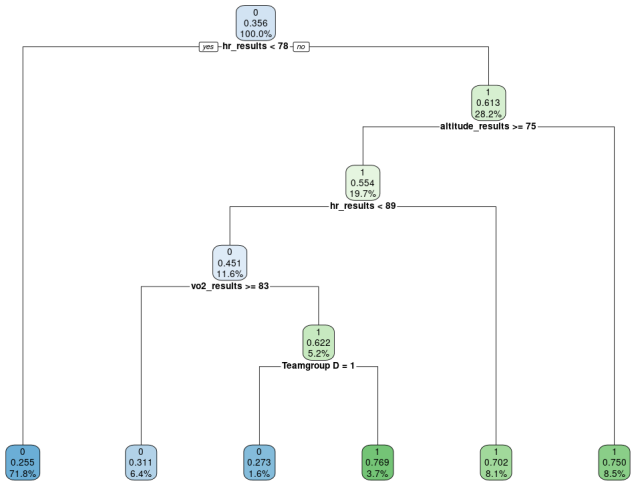

Fig. 11. Árvore de Decisão Winter Training Camp

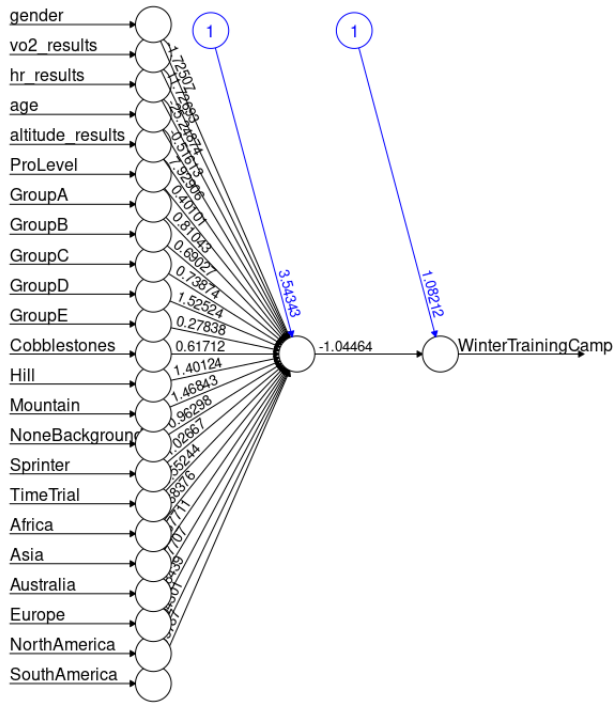

Fig. 12. Rede Neuronal Winter Training Camp

Já com a rede neuronal gerada obteve-se uma *accuracy* de aproximadamente **73.56%**, tendo sido realizada uma aproximação dos valores previstos para o atributo Winter Training Camp para os números inteiros mais próximos (i.e. 0 ou 1).

	Média	Desvio Padrão
Árvore de Decisão	0.69	0.02
Rede Neuronal	0.72	0.03

Através da análise dos valores obtidos para a média e desvio padrãos dos dois modelos conseguimos concluir que a Rede Neuronal apresenta uma maior *accuracy* mas a Árvore de Decião é mais consistente.

Recorreu-se a testes de hipótese para a avaliação das diferenças entre todos os valores de *accuracy* calculados anteriormente, concluindo que existem diferenças significativas no desempenho dos dois melhores modelos obtidos anteriormente, sendo o modelo Rede Neuronal o melhor.

Para não tornar exaustiva a comparação dos 2 modelos, resumiu-se os valores obtidos para cada critério (Accuracy, Sensitivity, Specificity e F1) na tabela seguinte:

	Árvore de Decisão	Rede Neuronal
Accuracy	67.8%	73.56%
Sensitivity	0.85	0.81
Specificity	0.33	0.59
F1	0.78	0.80

Atributo Gender

Com a finalidade de estudar a capacidade preditiva relativamente ao atributo Gender, foram obtidos 2 métodos de classificação, Rede Neuronal e K-vizinhos-mais-próximos. Inicialmente, recorreu-se a uma seed para que a reprodutibilidade dos resultados seja garantida.

Rede Neuronal

O modelo para a rede neuronal foi criado com uma arquitetura com 1 nó interno.

A *accuracy* do modelo gerado obteve o resultado satisfatório de **87.80%** tendo sido realizada uma aproximação dos valores previstos para o atributo Gender para os números inteiros mais proximos (i.e. 0 ou 1).

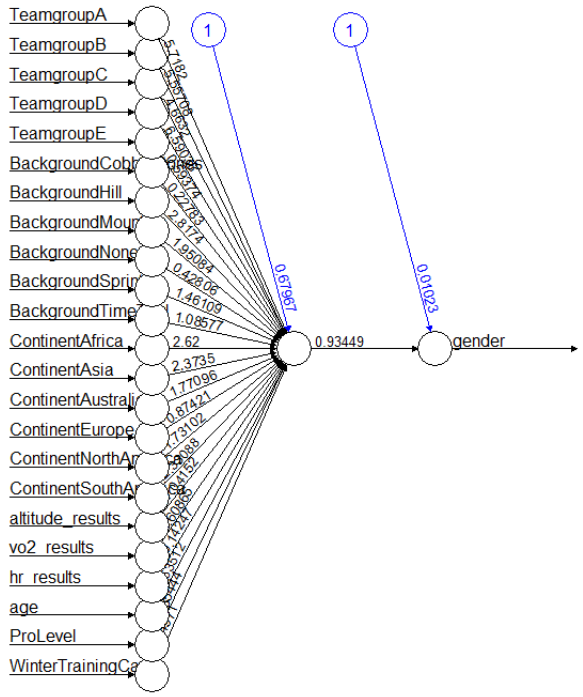

Fig. 13. Rede Neuronal Gender

K-vizinhos-mais-próximos

No modelo k-vizinhos mais próximos, obteve-se o K que maximizava a accuracy, 13, a que correspondia a accuracy máxima de **99.66%**.

Depois de analisados os 2 modelos conclui-se que o modelo com melhor capacidade preditiva é o K-vizinhos-mais-próximos. Posteriormente, foi utilizado o metodo k-fold cross validation para obter uma média e um desvio padrão da *accuracy* da previsão do atributo Gender, para estes dois modelos.

	Média	Desvio Padrão
Rede Neuronal	0.8804	0.0153
KNN	0.9895	0.0379

Foi, igualmente, realizado testes de hipótese para avaliar as diferenças entre todos os valores de *accuracy*calculados anteriormente, e conclui-se que existem diferencas significativas no desempenho dos modelos obtidos, sendo o modelo K-vizinhos-mais-próximos o melhor.

Na tabela seguinte, são apresentados os valores obtidos para cada métrica (Accuracy, Sensivity, Specificity e F1) em cada modelo.

	Rede Neuronal	KNN
Accuracy	87.8%	99.66%
Sensitivity	0.8853	1
Specificity	0.8695	0.9927
F1	0.8853	0.9968

Ainda que a diferença de desempenho entre os dois modelos não seja muito elevada, pode-se inferir a partir do teste de hipótese realizado que é significativa, em favorecimento do K-vizinhos-mais-próximos.

Conclusões

O presente trabalho constituiu uma fonte de aprendizagem valiosa aos autores dos diferentes algoritmos presentes na área de Machine Learning.

O tratamento de dados foi realizado adequadamente aos problemas propostos. Incitou ao espírito crítico dos autores ao propor que fossem criados, testados e comparados diversos modelos para a análise preditiva das variáveis pretendidas. A avaliação dos mesmos foi fundamentada através do estudo das métricas e tratamento de dados adequado aos problemas propostos, o que possibilitou compreender as principais diferenças e implicações na utilização dos diferentes modelos.

Nos problemas de regressão, salienta-se que foram feitos esforços para a obtenção dos modelos mais precisos para a predição das variáveis alvo respetivas. Após a obtenção da Regressão Linear Múltipla obtida para prever a variável altitude_results, foram analisados quais as variáveis que não eram estatísticamente significativas, tendo posteriormente sido criado um outro modelo sem essas mesmas variáveis, para melhores resultados. Adicionalmente, para todos os modelos de Rede Neuronal, foram analisadas diferentes arquiteturas de forma a escolher aquela que obtém um melhor desempenho. Desta forma, obtivemos valores de MAE e RMSE consistentemente satisfatórios. A título de exemplo, para prever o atributo vo2_results, a Regressão linear múltipla, Árvore de Regressão e Rede neuronal apresentaram valores de MAE inferiores a 5 e de RMSE entre os valores de 4 e 5.

Nos problemas de classificação, é de salientar que o algoritmo KNN-vizinhos-mais-próximos apresentou um desempenho elevado para a predição de ambas variáveis alvo ProLevel e Gender, com uma accuracy de 96.95% e de 99.66% respetivamente. Esses valores são ainda posteriormente reforçados pela média da taxa de acerto e desvio padrões obtidos através do método k-fold cross validation (valores de média de 0.97 e 0.98 e de desvio padrão de 0.01294 e 0.0379) e nas matrizes de confusão respetivas, com valores elevados em todos as métricas. Por outro lado, a Árvore de Decisão modelada para prever os atributos ProLevel e WinterTrainingCamp apresentou alguma inconsistência e variabilidade de resultados nas diferentes métricas das matrizes de confusão respetivas, ainda que com uma accuracy semelhante em relação a ambas as variáveis, rondando os 69%. Já a Rede Neuronal revelou resultados mais satisfatórios nas matrizes de confusão em relação aos atributos Gender e WinterTrainingCamp.

Em suma, foram dadas análises de dados apropriadas e essenciais ao problema proposto.

References

[1] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[2] C. Zhang and Y. Lu, "Study on artificial intelligence: The state of the art and future prospects," *Journal of Industrial Information Integration*, vol. 23, p. 100224, 2021.

[3] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[4] É. Fontana, "Introdução aos algoritmos de aprendizagem supervisionada," *Departamento de Engenharia Química, Universidade Federal do Paraná*, 2020.

[5] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[6] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[7] V. Sharma, S. Rai, and A. Dev, "A comprehensive study of artificial neural networks," *International Journal of Advanced research in computer science and software engineering*, vol. 2, no. 10, 2012.

[8] B. J. Wythoff, "Backpropagation neural networks: a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 2, pp. 115–155, 1993.

[9] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pp. 986–996, Springer, 2003.

[10] D. Berrar, "Cross-validation.," 2019.