Artigo Científico

André Teixeira
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Matosinhos, Portugal
1190384@isep.ipp.pt

Ivo Oliveira
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Matosinhos, Portugal
1190679@isep.ipp.pt

Resumo - Este artigo foi desenvolvido para a disciplina de Análise de Dados (ANADI) do curso de Engenharia Informática no Instituto Superior de Engenharia do Porto (ISEP). O objetivo principal foi aplicar os conhecimentos adquiridos ao longo do segundo semestre. Este artigo fornece uma discussão abrangente e interpretação dos tópicos abordados, acompanhados por uma infinidade de diagramas estatísticos para facilitar a resolução de questionamentos. São apresentadas explicações detalhadas e exemplos práticos para tópicos como regressão linear, árvores de decisão, redes neuronais e vizinhos mais próximos (*K-Nearest Neighbours*).

Palavras-chave: algoritmos, regressão, previsão, ciclista, classificação

I. Introdução

A utilização de algoritmos tem sido uma prática duradoura no mundo. Através dos algoritmos, temos sido capazes de prever eventos e resultados, o que tem sido crucial para a sobrevivência de nossa espécie e desenvolvimentos futuros. As máquinas, com seu poder computacional incrivelmente rápido, tornaram-se inestimáveis nesse sentido, capazes de processar e calcular milhões de resultados em frações de segundo.

Neste estudo, o foco foi recolher informações de ciclistas profissionais, especialmente em relação ao seu treino na pré-temporada. O conjunto de dados recolhido abrangia vários elementos relacionados ao regime de treino dos ciclistas. Para estabelecer potenciais relações entre os dados recolhidos e as métricas de desempenho que indicam o desempenho dos atletas, vários algoritmos foram utilizados: regressão linear, árvores de decisão, redes neuronais e vizinhos mais próximos (*K-Nearest Neighbours*). Esses algoritmos foram abordados extensivamente durante o semestre, e uma breve introdução contextual a cada algoritmo será fornecida para esclarecer quaisquer dúvidas iniciais. Posteriormente, eles foram aplicados à tarefa em questão para classificá-los e compará-los com base na qualidade dos resultados obtidos.

Para concluir o artigo, será apresentado um resumo dos resultados obtidos e das principais conclusões, proporcionando uma visão abrangente do estudo.

II. Estado de Arte

A. Regressão Linear

A Regressão Linear é um tipo de análise preditiva utilizada para prever o valor de uma variável (variável dependente) com base no valor de outra variável (variável independente) [1]. Ela estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes, que melhor preveem o valor da variável dependente [2].

B. Árvores de decisão

Ao contrário dos testes paramétricos, os testes não paramétricos exigem menos pressuposições em relação à população. O que define um teste não paramétrico é o fato de serem métodos de análise estatística que não exigem que uma distribuição atenda às suposições necessárias para serem analisadas. Se os dados não seguem uma distribuição normal, então devem ser utilizados testes não paramétricos [3]. Esses testes servem como uma alternativa aos testes paramétricos, como Anova e testes t. Alguns exemplos de testes não paramétricos são o Teste U de Mann-Whitney, Teste de Wilcoxon para amostras pareadas, Teste de Kruskal-Wallis, Teste de Friedman e Teste de Kolmogorov-Smirnov [3].

C. Redes Neuronais

As redes neuronais são uma série de algoritmos inspirados e que imitam a forma como o cérebro humano opera. Interconectadas por meio de nós, elas têm a capacidade de aprender a executar com sucesso tarefas que antes eram desconhecidas, solucionando problemas nos campos de aprendizagem de máquina e aprendizagem profunda [4]. A rede opera em camadas de nós, todos os quais se conectam entre si e partilham informações e dados [4].

D. Vizinho mais próximo (K-Nearest Neighbours)

O k-vizinhos mais próximos é um algoritmo usado para realizar previsões com base na identificação de semelhanças nas informações recebidas [5]. Não é paramétrico, o que significa que o seu modelo estrutural é determinado pelas informações utilizadas, e é por esse motivo que o KNN é um dos algoritmos mais utilizados nos tempos modernos [5].

III. Regressão

A. Preparação

Para ter acesso às informações obtidas, foi necessário carregar os dados, que estavam no formato CSV.

B. Visualização de Dados

Para obter uma melhor visualização das informações obtidas, foi solicitada a dimensão do conjunto de dados que continha as informações. Foi confirmado que o conjunto de dados possuía 1000 linhas e 11 colunas. Em seguida, as 11 colunas serão listadas:

- **ID**: numeração sequencial das informações para cada ciclista
- **gender**: o género de cada ciclista (masculino ou feminino)
- **Team**: o grupo ao qual o ciclista pertencia (do grupo A ao E)
- **Background:** o tipo de perfil do ciclista (Calçada, Colina, Montanha, Nenhum, Sprinter, Contrarrelógio)

- **Pro.level**: o nível de profissionalismo do ciclista (World Tour, Continental)
- **Winter.Training.Camp**: indicação se o ciclista concluiu ou não (concluído ou nenhum)
- **altitude_results**: resultado do treino em altitude
- **vo2_results** resultado do teste de volume máximo de oxigénio
- **hr_results**: resultado do teste de frequência cardíaca
- **dob**: Data de nascimento
- **Continent**: Continente de origem do ciclista

Além disso, foi solicitado um resumo dos dados do conjunto de dados, para obter uma primeira visão estatística desses dados. Nessa análise, foram observados aspectos como o valor mínimo, o primeiro quartil, a mediana, a média, o terceiro quartil e o valor máximo. Esse resumo pôde ser realizado apenas para as colunas cujos dados eram numéricos. Os resultados estão representados em Resultados 1.

```
       ID           gender              Team           Background         Pro.level         Winter.Training.Camp
 Min.   :  0.0   Length:1000        Length:1000        Length:1000        Length:1000        Length:1000
 1st Qu.:249.8   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :499.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :499.5
 3rd Qu.:749.2
 Max.   :999.0
 altitude_results  vo2_results       hr_results         dob             Continent
 Min.   : 24.00   Min.   : 21.00   Min.   : 17.00   Length:1000        Length:1000
 1st Qu.: 57.00   1st Qu.: 60.00   1st Qu.: 58.00   Class :character   Class :character
 Median : 68.00   Median : 70.00   Median : 69.00   Mode  :character   Mode  :character
 Mean   : 66.75   Mean   : 69.75   Mean   : 68.57
 3rd Qu.: 77.00   3rd Qu.: 80.00   3rd Qu.: 79.00
 Max.   :100.00   Max.   :100.00   Max.   :100.00
>
```

Resultados 1- Resumo das colunas numéricas

Posteriormente, foi observado que existia uma coluna chamada "dob" (sigla para "Data de Nascimento"), cujos dados estavam no formato de data, tornando-os irrelevantes para o estudo. Portanto, a fim de torná-los relevantes, eles foram convertidos para a idade de cada ciclista, colocando-os em uma nova coluna chamada "Age". O processo usado nessa conversão começou por formatar as datas na coluna dob, que estavam em formato de texto, para o tipo de data, com o formato "%Y-%m-%d", onde Y representa o campo do ano, m para o mês e d para o dia. Finalmente, para converter a data de nascimento do ciclista em sua idade atual, foi calculada a diferença entre a data atual e a data na coluna e dividida por 365,25. Optou-se por dividir por esse número e não por 365, pois a cada 4 anos há um ano com um dia a mais, portanto, ao adicionar 0,25 dias, aumentamos a precisão do cálculo.

C. Análise de Dados

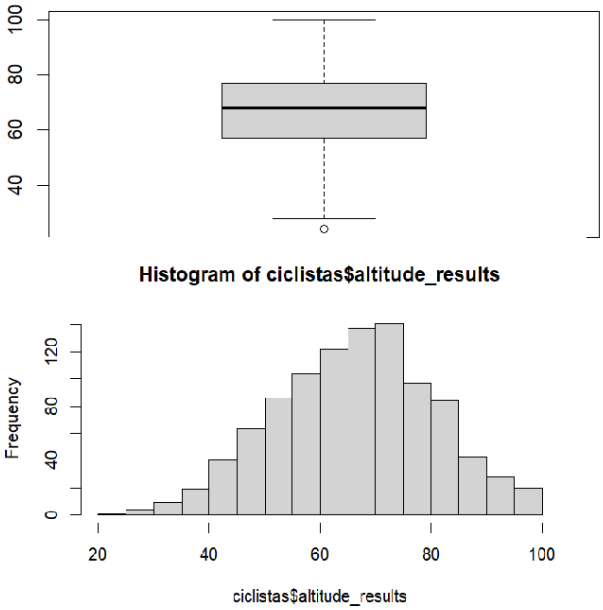
Para analisar estatisticamente todas as colunas numéricas, além dos aspectos obtidos ao resumir essas colunas, foram utilizadas Medidas de Variabilidade/Dispersão e Medidas de Forma.

Para iniciar, nas Medidas de Variabilidade/Dispersão, foram calculados a amplitude geral, a amplitude interquartil (AIQ), a variância e o desvio padrão dos dados de cada coluna numérica. Os resultados estão apresentados na Tabela 1.

Tabela 1- Medidas de Variabilidade/dispersão

	Ampltiude	IQR	Variância	Desvio Padrão
altitude_results	76	20	203,68	14,27164
vo2_results	79	20	197,024	14,03652
hr_results	83	21	216,433	14,71167
Age	21	10	34,4697	5,87109

Foram utilizados gráficos de boxplot, histogramas, skewness e kurtosis. Os gráficos serão apresentados em forma de mosaico, sendo o gráfico de boxplot no topo e o histograma na parte inferior.
Abaixo dos gráficos, também serão apresentados skewness e kurtosis associadas




```
> skewness(ciclistas$altitude_results) #-0.1151302
[1] -0.1151302
> kurtosis(ciclistas$altitude_results) #-0.3417119
[1] -0.3417119
```

Figura 1- Medidas de forma dos Altituds_results

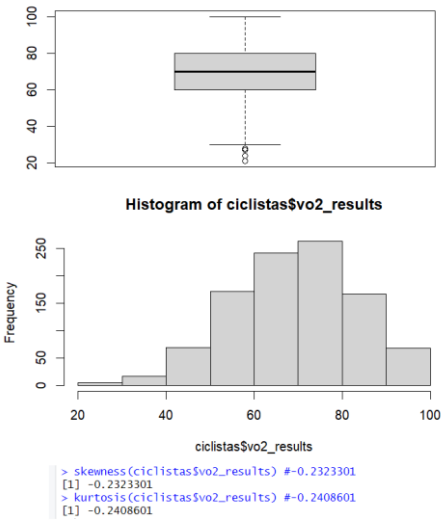


Figura 2- Medidas de forma de vo2_results

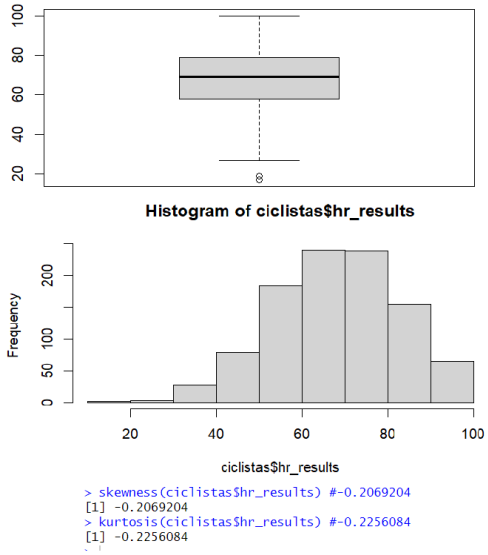


Figure 3- Medidas de forma de hr_results

Figura 4- Medidas de forma de Age

As Medidas de Forma também foram aplicadas às colunas com dados não numéricos, na forma de gráficos de pizza.

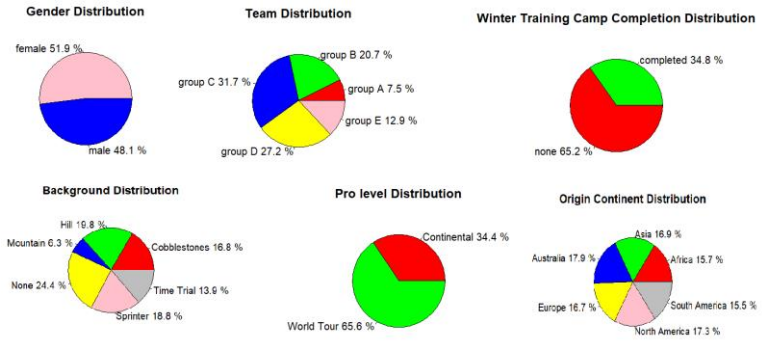


Figura 5- Medidas de forma com colunas não numericas

D. Processamento de Dados

A verificação da existência ou não de dados inválidos, como NA, foi realizada. Verificou-se que não havia dados inválidos. Se houvesse algum dado desse tipo, a linha seria excluída.

Dito isso, procedemos à identificação dos valores discrepantes (outliers) das colunas numéricas, utilizando o AIQ. Vale ressaltar que esses valores também podem ser identificados pelos gráficos de boxplot.

Usando o AIQ, são considerados outliers os valores que estão fora do intervalo fechado entre o primeiro quartil e o terceiro quartil do conjunto de dados.

Outliers of:

- altitude_results →24
- vo2_results → 21, 24, 27, 28
- hr_results → 17, 19
- Age → none

Durante essa identificação, verificou-se que as colunas ID e dob não eram mais relevantes para o estudo, portanto, foram eliminadas.

E. Correlação

Na preparação para criar o diagrama de correlação, procedemos à normalização dos dados, convertendo-os em dados numéricos e iniciando-os em 0.

Neste ponto, já foi possível criar o diagrama de correlação dos dados, representado pela Figura 6.

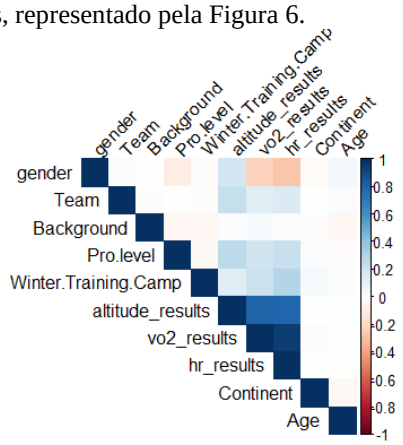


Figura 6- Diagrama Correlação

Podemos observar uma correlação muito forte entre os resultados de vo2_results e hr_results, e correlações fortes entre altitude_results e vo2_results e altitude_results e hr_results. As variáveis Team e Background apresentam correlações muito fracas com os demais aspectos, o que significa que os resultados de cada ciclista não dependem nem do seu nível profissional nem da equipe a que pertencem.

F. Correlação

Foi solicitado um modelo de regressão linear para determinar os resultados do teste de altitude utilizando o atributo da frequência cardíaca. Nesse modelo, a equipa procurou garantir que a variável dependente fosse justificável por uma ou várias variáveis independentes.

O primeiro passo foi separar os valores em casos de teste e treino. Os primeiros foram usados para construir o modelo, enquanto os últimos foram usados para fazer comparações com o modelo, permitindo uma avaliação precisa do erro. Dessa forma, utilizamos a função lm para construir o modelo com os dados de treino. A função resultante nos fornece os valores precisos para determinar a função de regressão linear, que neste caso é Altitude_results = 14.625 + 0.761 * hr_results.

Ao definir uma função linear, a equipe procedeu à criação de um gráfico de dispersão para visualizar os dados. A Figura X mostra o diagrama resultante.

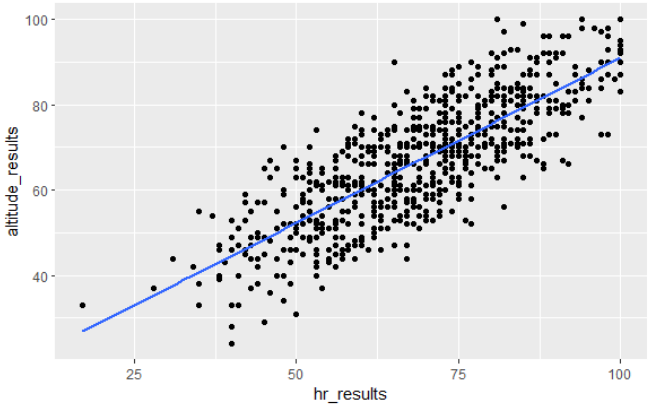


Figura 7 - Dispersão

Nesta última parte, tentamos obter melhores resultados com um modelo mais complexo, mais precisamente uma função polinomial com a adição de variáveis independentes mais significativas para os resultados de altitude. Dessa forma, utilizamos a função lm para construir o modelo com os dados de treinamento. A função resultante é Altitude_results = 9.93 + 0.518 * hr_results + 0.383 * vo2_results - 12.4 * gender + 0.048 * Background + 3.03 * ProLevel - 0.038 * Age.
Também calculamos o erro absoluto médio (MAE) e a raiz quadrada do erro médio (RMSE) para comparar com os erros da função linear. Ao comparar os valores do MAE (6.772438 / 4.693006) e do RMSE (8.28004 / 5.762637), notamos que os resultados da segunda função são significativamente menores e, portanto, esse modelo é relativamente melhor que o anterior.

G. Previsão do volume máximo de oxigénio

1) Regressão Linear Múltipla

Para realizar a regressão linear, foram utilizados os dados normalizados, calculados a partir dos dados originais obtidos.

Primeiramente, foram criados casos de treino e casos de teste de acordo com o método de Holdout (70% para os casos de treino e 30% para os casos de teste). Os casos de treino são utilizados para construir o modelo desejado e os casos de teste são usados para comparar o modelo e avaliá-lo em termos de precisão e erro.

Em segundo lugar, para construir o modelo, foi utilizada a função 'lm', utilizando os casos de treino e a fórmula com todos os atributos.

A equação resultante da regressão linear múltipla foi: 0.01768 + 0.01184 * Gender – 0.02236 * Team + 0.00180 * Background – 0.01444 * Pro.level – 0.02137 * Winter.Training.Camp + 0.17542 * altitude_results + 0.83417 * hr_results + 0.00309 * Continent + 0.00647 * Age

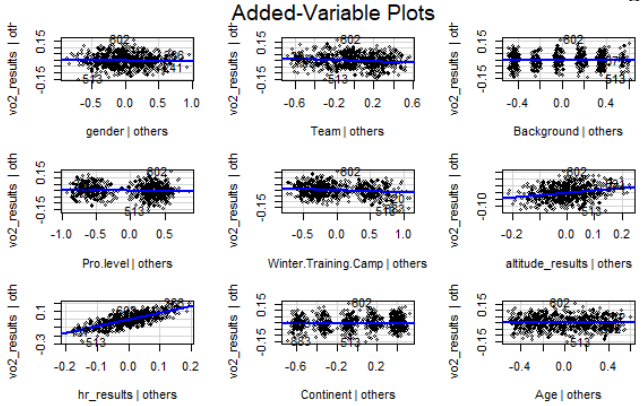


Figura 8 – Regressão Linear Múltipla

2) Árvore de regressão

Os dados (normalizados) utilizados são os mesmos do modelo anterior para fazer uma comparação justa, incluindo os casos de teste e treino e os atributos usados no modelo.

Para criar a árvore de regressão, foi utilizada a função 'rpart', utilizando os casos de treino e a mesma fórmula para desenvolver o modelo.

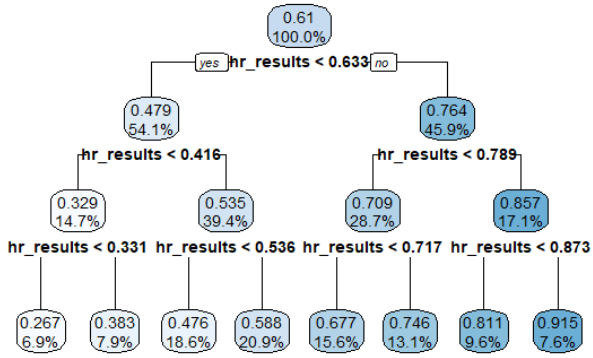


Figura 9 – Árvore de Regressão

3) Rede Neuronal

Os mesmos procedimentos foram realizados para construir a Rede Neural, relativos aos casos de treino e teste, dados utilizados e fórmula.

Para criar a Rede Neuronal, foi utilizada a função 'neuralnet'. O número de nós utilizados é equivalente a 1, apenas para facilitar a legibilidade dos dados.

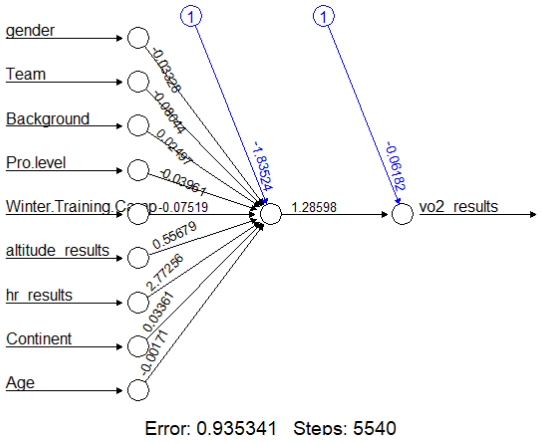


Figura 10 – Rede Neuronal

4) Comparação dos modelos

Após calcular todos os modelos, foi calculado o erro absoluto médio (MAE) e o erro quadrático médio (RMSE) para cada modelo.

	Regressão Linear Múltipla	Árvore Decisão	Rede Neuronal
MAE	0.0447367	0.0510819	0.0458417
RMSE	0.0548960	0.0656089	0.0557540

Ao comparar os resultados, a conclusão é que os dois melhores modelos são a regressão linear múltipla e a rede neuronal.
Para verificar se existem diferenças significativas entre esses dois melhores modelos, primeiro precisamos saber se eles seguem uma distribuição normal. Para isso, foi realizado o teste de Lilliefors, pois trata-se de uma amostra grande. Como o valor para ambos os valores de p (0,9125 para a regressão linear múltipla e 0,4845 para a rede neuronal) é maior do que o nível de significância (5%), a hipótese nula é aceita e conclui-se que ambas as distribuições são normais.

Portanto, é seguro utilizar o 't.test' para comparar os dois modelos. O resultado do valor de p é 0.9027 , portanto, a hipótese nula não é rejeitada e conclui-se que não há diferenças significativas entre os dois modelos.

IV. Classificação

A. *Previsão de Pro_Level*

1) Capacidade Previsão

Para estudar a capacidade preditiva do atributo "Pro_Level", foram utilizados os métodos "Árvores de Decisão", "Rede Neuronal" e "K-vizinhos mais próximos". Primeiramente, foi estudada a precisão de cada um deles.

O primeiro método mencionado, representado pela Figura 11, resultou numa precisão de 1,33%

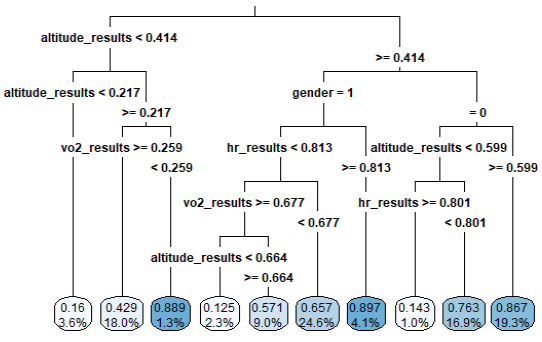


Figura 11- Árvore de Decisão

O segundo método mencionado, representado pela Figura 12, resultou numa precisão de 6,33%

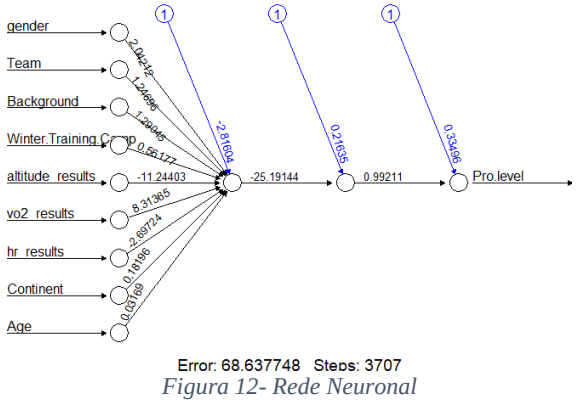


Figura 12- Rede Neuronal

O último método mencionado, representado pela Figura 13, resultou numa precisão de 4.666667%

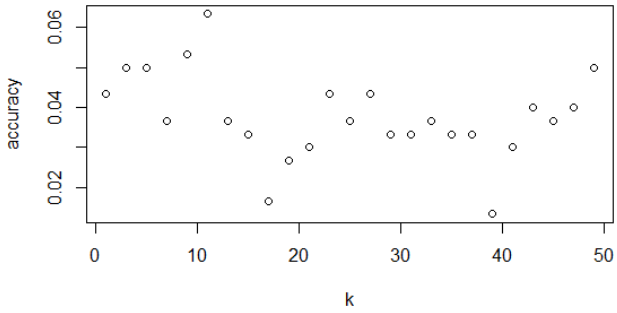


Figura 13- K-Vizinhos mais próximos

Com base nesses resultados, pode-se concluir que os melhores modelos obtidos foram a Rede Neuronal e os K-vizinhos mais próximos. Dito isto, podemos usar o método k-fold para obter a média e o desvio padrão da previsão do atributo Pro_Level.

Através deste estudo, foi concluído que o método de Redes Neuronais teve uma média de 0.058331529 e um desvio padrão de 0.019285106. Por outro lado, o método de K-vizinhos mais próximos teve uma média de 0.006102975 e um desvio padrão de 0.005264771.

Dos três modelos, um é conhecido por ter uma curva de aprendizagem, conhecido como "Lazy Learning".

O modelo conhecido como "k-vizinhos mais próximos" (K-NN) ou "Lazy learning" é conhecido pelo seu método de aprendizagem "Lazy learning". Devido à sua abordagem de treino, o K-NN é considerado um algoritmo de aprendizagem preguiçoso (Lazy learning).

No K-NN, o modelo não constrói explicitamente uma representação do conjunto de treino durante a mesma.

Em vez disso, ele armazena todos os seus dados de treino em bancos de dados de treino e utiliza-os para fazer previsões posteriormente. O algoritmo não adiciona características ou aprende um modelo específico durante a fase de treino. Em vez disso, ele repete todo o regime de treino na forma bruta.

Quando um novo ponto de dados precisa ser previsto ou classificado, o algoritmo K-NN calcula a distância entre esse ponto de dados e os outros pontos de dados no conjunto de treino. Em seguida, ele escolhe os k pontos mais próximos (daí o nome "K-NN") e usa as classes ou valores desses pontos para criar sua previsão.

Implicações de K-NN "Lazy Learning":

1. Custo computacional durante a fase de inferência: O k-NN precisa calcular as distâncias entre o novo ponto de dados e cada ponto de dados no conjunto de treino combinado durante a fase de inferência. Isso pode ser computacionalmente caro, especialmente para grandes coleções de dados.
2. Armazenamento completo do conjunto de treino: O k-NN requer o armazenamento completo do conjunto de treino, o que pode ser ineficiente em termos de uso de memória ao lidar com conjuntos de dados grandes.
3. Decisões locais: O k-NN toma decisões com base nos vizinhos mais próximos, o que significa que ele pode aprender padrões locais, mas pode não capturar relacionamentos globais complexos nos dados.
4. Sensibilidade à dimensionalidade: O desempenho do k-NN pode ser prejudicado pela distorção da dimensionalidade. À medida que as dimensões dos dados aumentam, o espaço de busca por vizinhos próximos também aumenta, o que pode resultar em um desempenho reduzido.
5. Sensibilidade à escala e ao ruído: O algoritmo k-NN é sensível à escala dos recursos, portanto, é necessário normalizar os dados antes de usar o algoritmo. Além disso, o k-NN também pode ser sensível a ruídos ou pontos inconsistentes com os dados.

Em resumo, o k-NN como um modelo de aprendizagem preguiçoso oferece uma abordagem simples e intuitiva para classificação e regressão, mas tem implicações em termos de armazenamento de dados, custo computacional e sensibilidade a fatores como dimensionalidade, escala e ruído.

Posteriormente, verificamos se os dois modelos eram significativamente diferentes, com um nível de significância de 5%, em que H0 é "sem diferenças significativas" e H1 é "existem diferenças significativas".

Após o teste, como p = 0.7291374 e, portanto, é maior do que 0,05, não podemos rejeitar a hipótese nula. Não há evidências estatísticas, com um nível de significância de 5%, que comprovem que existem diferenças significativas entre os dois modelos.

2) Análise de perfomance

Para comparer os resultados utilizamos os critérios de Accuracy, Sensitivity; Specificity e F1. Os resultos estão na tabela 2.

Tabela 2- Comparação de resultados

	Rede Neuronal	K-NN
Accuracy	0.3533333	0.0166666 7
Sensitivity	0.6343612	0.3333333
Specificity	0.7484557	0.2
F1	0.6743612	0.25

De acordo com os resultados, podemos concluir que o modelo apresentado que tem melhor performance de acordo com os critérios é a Rede Neuronal.

B. Previsão do Winter Training Camp

1) Capacidade preditiva

Para estudar a capacidade do atributo Winter_Training_Camp usamos métodos de "Árvore de Decisão" e 'Rede Neuronal'

Primeiramente, foi realizado o modelo de árvore de decisão. Os dados utilizados foram os dados originais normalizados, e os casos de treino (70%) e teste (30%) foram obtidos a partir desses dados normalizados.
Como se trata de uma árvore de decisão, o método escolhido para o modelo foi o 'Class' (classe), e todos os atributos fazem parte da fórmula utilizada.

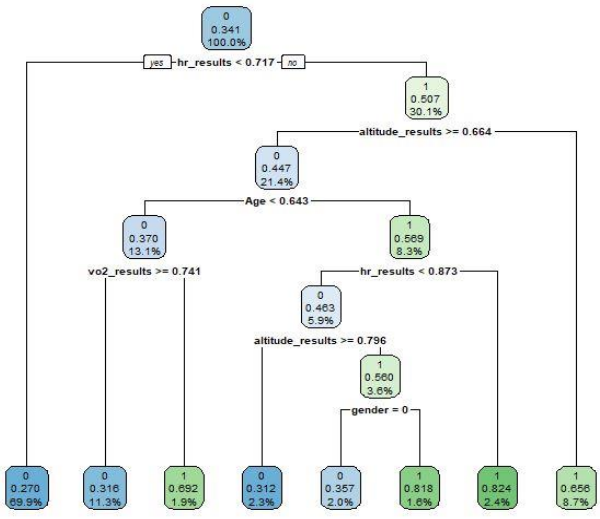


Figura 14 – Árvore de decisão

A precisão deste modelo foi de 0.6966667.

Para a 'Rede Neuronal', foi adotada a mesma abordagem em termos de dados utilizados, fórmula e casos de treino e teste.
O número de nós utilizados foi 1 devido à legibilidade dos dados.

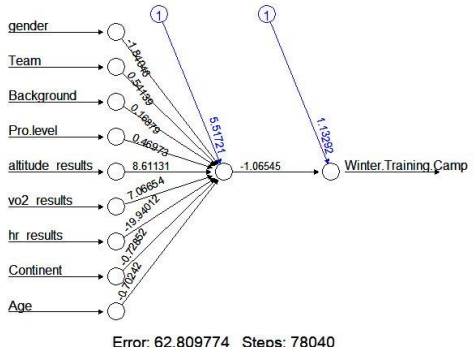


Figura 15 – Rede Neuronal

A precisão para este modelo foi de 0.7066667.

2) K-Fold Cross Validation

Os modelos utilizados foram os mesmos do tópico anterior.

Para descobrir a média e o desvio padrão da taxa de acerto da previsão do atributo Winter_Training_Camp.

A função 'for' realiza um loop 10 vezes e, a cada iteração, insere na matriz os valores da precisão dos modelos e suas previsões.

No final, os resultados foram obtidos utilizando a função 'apply', com a matriz e os atributos 'mean' (média) e 'sd' (desvio padrão).

Tabela 4 - Comparação de resultados

	Árvore Decisão	Rede Neuronal
Média	0.7015413	0.7323471
Desvio padrão	0.04000683	0.0398576

3) Teste de Hipótese

Foi realizado um teste de hipótese para verificar se existem diferenças significativas entre os dois modelos, com um nível de significância de 5%.

Primeiramente, foi realizado um teste de Lilliefors para verificar se a distribuição é normal.

Como o valor p foi menor que o nível de significância, foi realizado o teste de Wilcoxson. O resultado do valor p (0.2176) foi maior que o nível de significância, portanto, assume-se que não há evidências estatísticas de que os dois modelos apresentam diferenças significativas.

Tabela 4 - Comparação de resultados

	Árvore Decisão	Rede Neuronal
Accuracy	0.7066667	0.7233333
Sensitivity	0.7	0.7268722
Specificity	0.9162304	0.8638743
F1	0.7936508	0.7894737

4) Comparação de modelos

Ao comparar os dois modelos utilizando uma matriz de confusão para obter os valores de Accuracy, Sensitivity, Specificity e F1, podemos analisar os resultados obtidos:

Ao comparar os dois modelos, a Accuracy, Sensitivity foram maiores na rede neuronal, o que significa que ela é mais capaz de prever corretamente a maioria das amostras (Accuracy), é melhor em identificar corretamente as amostras da classe positiva (Sensitivity). A Specificity foi maior no modelo de árvore de decisão, o que significa que esse modelo é melhor em identificar corretamente as amostras da classe negativa.

Tabela 5 - Comparação de resultados

	Rede Neuronal	K-NN
Taxa de Acerto Média	0.0	0.634
Desvio padrão da Taxa de Acerto	0.0	0.04087923

5) Ao calcular os valores da taxa de acerto média e desvio padrão da taxa de acerto da previsão do atributo Gender, atingimos os resultados acima indicados, os valores da rede neuronal ficam logicamente a desejar, enquanto os valores relativos ao KNN indicam que em média, o modelo classificou corretamente 63.4% das amostras de teste e apontam uma variação de aproximadamente 0.04087923 na taxa de acerto entre as execuções. Após isso, feito um teste de Wilcoxon pareado, atingindo o valor p = 0.005825024, que por sua vez é menor que 0.05, afirmando assim que existem de facto diferenças significativas.

6) Analisando os resultados do último exercício, observamos que os resultados sugerem que a Rede Neural obteve uma precisão maior, uma sensibilidade superior e um valor de F1 mais elevado em comparação com o modelo KNN. No entanto, a especificidade da Rede Neural foi menor do que a do modelo KNN, como podemos ver representado abaixo.

Tabela 6 – Comparação de resultados

	Árvore Decisão	Rede Neuronal
Accuracy	0.7066667	0.7233333
Sensitivity	0.7	0.7268722
Specificity	0.9162304	0.8638743
F1	0.7936508	0.7894737

Conclusão

Este artigo científico explora a aplicação de algoritmos de aprendizagem de máquina para analisar dados relacionados aos treinos de pré-temporada de ciclistas profissionais. O estudo utilizou diversos algoritmos, incluindo redes neuronais, árvores de decisão, árvores de regressão e k-vizinhos mais próximos. Destaca-se que uma análise de regressão linear simples revelou um impacto significativo do atributo "Altitude_results" no atributo " hr_results". Além disso, uma análise de regressão linear múltipla identificou o atributo " hr_results" como o fator mais influente que afeta o nível máximo de oxigênio. Ao utilizar esses modelos, métodos e testes, o intérprete dos resultados obtém insights valiosos para projetar programas de treino melhores e mais direcionados, levando em consideração até mesmo os menores fatores para aprimorar o desempenho do ciclista.

Referências

[1] IBM, "About Linear Regression," *www.ibm.com*, 2023. https://www.ibm.com/topics/linear-regression (acedido Jun. 18, 2023).

[2] Statistics Solutions, "What is Linear Regression?," *www.statisticssolutions.com*, 2023. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is- linear-regression/ (Acedido Jun. 18, 2023).

[3] A. M. Madureira, "Decision Trees," 2022.

[4] IBM, "What are Neural Networks?," *www.ibm.com*,2023. https://www.ibm.com/topics/neural-networks (acedido Jun. 26, 2023).

[5] IBM, "What is the k-nearest neighbors algorithm?," *www.ibm.com*, 2023. https://www.ibm.com/topics/knn (acedido Jun. 26,2023).