Análise de Dados em Informática

José Maia  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1191419@isep.ipp.pt

Lourenço Melo  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1190811@isep.ipp.pt

Artur Muiria  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1161274@isep.ipp.pt

# Introdução

Este relatório descreve o segundo projeto prático realizado como parte da disciplina de Análise de Dados em Informática (ANADI), no curso de Engenharia Informática (LEI) do Instituto Superior de Engenharia do Porto.

O enunciado contém um ficheiro de dados. Foi feita a Análise Exploratória de Dados, Análise Inferencial, Correlação e Regressão para os dados fornecidos. Vai ser explicado como foram filtrados os dados e apresentadas as resoluções. Neste projeto foram estudados modelos como: regressão linear simples, regressão linear múltipla, árvores de decisão e regressão e redes neuronais. Estes modelos foram comparados e utilizados para analisar diversos dados, criando assim, uma analise profunda e detalhada acerca da performance e usabilidade dos mesmos.

# Metodologia de trabalho

Primeiramente, o grupo analisou o enunciado em conjunto e estudou os requerimentos de todas as alíneas. Durante a realização do trabalho, o grupo trabalhou e tirou dúvidas durante a aula e trabalhou em casa comunicando via "Discord". Mesmo tendo havido divisão de tarefas, houve entreajuda e participação de todos os elementos do grupo na totalidade do trabalho.

Para a resolução dos exercícios foi utilizada a linguagem R e a plataforma RStudio, para além desta ferramenta usamos toda a informação disponibilizada no moodle da unidade curricular. Finalmente, depois da resolução de todos os exercícios passamos à elaboração do artigo científico em grupo onde revimos todos os problemas e discutimos resultados.

# Objetivos

Os objetivos deste projeto prático são adquirir um entendimento aprofundado dos diferentes modelos de classificação e regressão abordados nesta disciplina, bem como aplicar os algoritmos de aprendizagem automática conhecidos. Pretende-se utilizar os algoritmos para realizar uma análise de dados precisa e, assim, resolver de forma correta todas as questões apresentadas no enunciado.

# Estado de Arte

***1 - Machine Learning***

No ramo da inteligência artificial conhecido como "machine learning", os investigadores trabalham para criar algoritmos e métodos que permitam aos computadores aprender, raciocinar e realizar tarefas sem terem de ser explicitamente programados. Os algoritmos de aprendizagem automática são criados para examinar grandes volumes de dados, detetar padrões e aprender com eles para fazer previsões ou conclusões, em vez de seguir instruções precisas.

Três componentes principais constituem o sistema de aprendizagem de um algoritmo de aprendizagem automática:

1) **Tomada de decisão**: Normalmente, as previsões ou classificações são efetuadas utilizando algoritmos de aprendizagem automática. O programa irá gerar uma estimativa dos padrões descobertos nos dados com base num conjunto de dados.

2) **Função de erro**: A previsão efetuada pelo modelo é avaliada utilizando a função de erro. Se estiverem disponíveis exemplos rotulados, esta função pode comparar e avaliar a exatidão do modelo.

3) **Processo de otimização do modelo**: Se o modelo se ajustar melhor ao conjunto de dados de treino, ajusta os seus pesos para reduzir a discrepância entre o exemplo conhecido e a estimativa do modelo. O algoritmo repete este processo de avaliação e otimização, atualizando autonomamente os pesos até atingir um nível máximo de precisão.

***2 - Árvores de Decisão***

Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas. Permite que um indivíduo ou organização compare possíveis ações com base em seus custos, probabilidades e benefícios. Podem pode ser usada tanto para conduzir diálogos informais quanto para mapear um algoritmo que prevê a melhor escolha, matematicamente. Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore. (O que é um diagrama de árvore de decisão?)

Em conclusão, as árvores de decisão são estruturas em forma de árvore que dividem os dados recursivamente em categorias ou previsões em função dos atributos. Proporcionam flexibilidade e capacidade no processamento de vários tipos de dados.

***3- Árvores de Regressão***

As árvores de regressão são em tudo idênticas às árvores de decisão analisadas na Secção 2. A diferença principal reside no facto das folhas das primeiras conterem previsões numéricas e não decisões.

A obtenção de árvores de regressão usando o R é conseguida usando exatamente a mesma função usada para as árvores de decisão. De facto, a função rpart vai obter uma árvore de regressão ou de decisão consoante o tipo da variável objetivo. Se esta for um fator, a função obtém uma árvore de decisão, se for uma variável numérica é obtida uma árvore de regressão. De resto toda a sintaxe é igual, embora a função possua parâmetros específicos para cada tipo de árvore. (Torgo, 2003)

***4 – Regressão Linear***

A regressão linear é uma técnica de modelação estatística utilizada para analisar a relação entre uma variável dependente e uma ou mais variáveis independentes. Pressupõe uma relação linear entre as variáveis e tem como objetivo encontrar a linha de melhor ajuste que representa essa relação.

Na regressão linear, a variável dependente é frequentemente referida como a variável-alvo ou a variável de resposta, enquanto as variáveis independentes são conhecidas como variáveis características. O objetivo é estimar os coeficientes da equação linear que minimiza a diferença entre os valores previstos e os valores reais da variável dependente.

***5 - K-vizinhos-mais-próximos***

O algoritmo k-vizinhos mais próximos (do inglês, k-Nearest Neighbors – kNN) funciona da seguinte forma: dada uma instância de teste *xq*, o algoritmo encontra os k vizinhos mais próximos de *xq* no conjunto de treino. Em seguida, a classe de *xq* é dada pela classe que ocorrer com maior frequência entre os k vizinhos. (Darmiton, 2020)

***6 – Redes Neuronais***

Redes neuronais são sistemas de computação com nós interconectados que funcionam como os neurônios do cérebro humano. Usando algoritmos, podem reconhecer padrões escondidos e correlações em dados brutos, agrupá-los e classificá-los, e – com o tempo – aprender e melhorar continuamente. (Redes Neurais, s.d.)

Uma rede neuronal simples inclui uma camada de entrada, outra de saída (ou alvo) e, entre elas, uma camada oculta. As camadas são conectadas através de nós e essas conexões formam uma "rede" – a rede neural – de nós interconectados.

# Análise e Discussão de Resultados

*Exercício 1:* Comece por carregar o ficheiro (“ciclismo.csv”) para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.

Começou-se por carregar o ficheiro “ciclismo.csv” que contém os dados necessários para a resolução do enunciado. O ficheiro possui dados acerca da performance de ciclistas profissionais, incluindo elementos relativos aos treinos no período de pré-temporada. Verificou-se a existência de 1000 observações e 13 variáveis. Cada uma das 1000 linhas representa uma instância ou registo de dados, enquanto as 13 colunas representam diferentes características ou variáveis associadas a cada instância.

Finalmente, utilizando a função “*summary*” conseguimos observar a alguns dados relevantes acerca da amostra tais como, a média, a mediana, mínimos e máximos.

*Exercicio 2:* Derive um novo atributo *Age* usando como valor do atributo dob.

Utilizando a coluna *dob* que representa a data de nascimento de cada ciclista, foi criado um atributo Age que apresenta a idade respetiva de cada um. Para este propósito utilizamos a seguinte função:



*Exercicio 3:* Analise os atributos do conjunto de dados mais significativos, usando gráficos, análises estatísticas e/ou outros métodos apropriados.

Inicialmente, começou-se por criar uma variável “*intervalo*\_*idades*” onde utilizando a coluna Age dividindo-se os ciclistas for faixa etária dos 0 aos 20, 21 aos 30 e 31 aos 40 anos. Assim, podemos utilizar esta coluna para relacionar com as outras variáveis da amostra, tornando os gráficos mais simples e percetíveis.

Primeiramente, criou-se um *boxplot* somente com os dados relativos ao teste de volume de oxigênio máximo, obtendo-se o seguinte gráfico:

Figura 2 - Boxplot relativo ao teste de volume de oxigênio máximo

Uma imagem com texto, captura de ecrã, diagrama, Retângulo

Descrição gerada automaticamente

De seguida, criou-se outro *boxplot* onde comparamos o teste de volume de oxigênio máximo em relação ao intervalo de idade e ao background do ciclista. O *boxplot* obtido foi o seguinte:

Figura 3 - Boxplot relativo ao teste de volume de oxigênio máximo em relação ao intervalo de idade e ao background do ciclistaUma imagem com texto, diagrama, captura de ecrã, Gráfico

Descrição gerada automaticamente

Criou-se outro *boxplot* igual ao anterior, mas com os dados relativos ao resultado do teste de frequência cardíaca.

Figura 4- Boxplot relativo ao teste de frequência cardíaca em relação ao intervalo de idade e ao background do ciclista.

Uma imagem com texto, diagrama, Gráfico, captura de ecrã

Descrição gerada automaticamente

Analisando estes gráficos podemos verificar a distribuição dos resultados dos testes por intervalo de idades e género, comparar o background dos ciclistas com os resultados nos testes e identificar possíveis outliers.

*Exercício 4:* Realize o pré-processamento dos dados:

a) *Faça a identificação de NA e limpe o dataSet, se aplicável.*

Procedeu-se à limpeza do dataSet e verificou-se que não apresentava nenhum NA.

b) *Identifique dados inconsistentes e outliers, se aplicável.*

Para proceder à identificação de outliers representou-se cada variável em função do treino de altitude. Para a variável gender obtivemos o seguinte *boxplot:*

Figura 5 - Boxplot do gender em função do treino de altitude.

Uma imagem com Retângulo, diagrama, captura de ecrã, Paralelo

Descrição gerada automaticamente

Verifica-se que a média de ciclistas do sexo masculino é superior e que existem dois outliers no campo dos ciclistas masculinos.

Outro exemplo é o gráfico relativo à equipa do ciclista.

Figura 6 - Boxplot da equipa do ciclista em função do treino em altitude.

Uma imagem com diagrama, Retângulo, texto, captura de ecrã

Descrição gerada automaticamente

Conclui-se que existem cinco grupos, a média do grupo E é superior ao resto e a existência de alguns outliers no grupo C, D e E.

c) *Implemente a seleção de atributos, se aplicável.* Calculou-se a importância dos atributos em função do vo2\_results e conclui-se que as variáveis winter\_camp, altitude\_results e age têm a maior importância.

d) *Implemente a normalização dos dados, se necessário.*

Primeiramente normalizou-se os valores das colunas “altitude\_results”, “hr\_results” e “vo2\_results”. De seguida, passaram-se os dados do tipo categóricos para tipo o numérico.

*Exercicio 5:* Crie um diagrama de correlação entre todos os atributos. Comente o que observa.

Criou-se um diagrama de correlação com todos os atributos incluindo o intervalo de idades e a idade.

Figura 7 - Diagrama de correlaçãoUma imagem com texto, captura de ecrã, número, diagrama

Descrição gerada automaticamente

Analisando o gráfico conclui-se que existe um grande grau de correlação entre o hr\_results e a altitude\_results, o hr\_results e o vo2\_results e o v02\_results e a atitude\_results. Verifica-se que ambos os testes e o treino de altitude possuem um elevado grau de correlação entre eles.

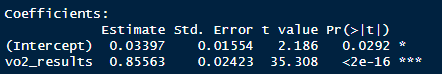
Para além destas correlações positivas, encontram-se fortes correlações negativas, onde quando uma variável aumenta a outra diminui, entre o hr\_results e género, vo2\_results e género e winter\_camp e hr\_results.

*Exercício 6:* Obtenha um modelo de regressão linear simples para determinar a variável “Altitude\_results” usando o valor relativo à componente dos resultados de frequência cardíaca (“hr\_results”).

Começou-se por separar os dados em conjunto de teste e conjunto de treino.

*a) Apresente a função linear resultante.*

Obteve-se os seguintes valores:



Concluímos que a função é:

AltitudeResults=0.03397+(vo2\_results\*0.85564)

b) *Visualize a reta correspondente ao modelo de regressão linear simples e o respetivo diagrama de dispersão.*

Figura 8 - Diagrama de dispersão com a reta de regressão linear simplesUma imagem com texto, diagrama, captura de ecrã, file

Descrição gerada automaticamente

Analisando o diagrama, verifica-se que a altitude\_results aumenta em proporcionalidade direta com o vo2\_results.

c) *Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) do modelo sobre os 30% casos de teste.*

O MAE e o RMSE são medidas comuns de desempenho usadas para avaliar a precisão de um modelo de regressão. O MAE representa a média das diferenças absolutas entre os valores previstos e os valores reais, enquanto o RMSE é a raiz quadrada da média dos erros ao quadrado. Quanto menor o valor do MAE e do RMSE, melhor o desempenho do modelo, pois indicam que as previsões estão mais próximas dos valores reais.

Obteve-se os seguintes valores:

**MAE**: 0.0836095174587336.

**RMSE**: 0.103535616413071.

*d) Teste se é possível obter resultados melhores utilizando um modelo mais complexo.*

Para este exercício, optou-se por obter os resultados utilizando o modelo de regressão linear múltipla. No modelo de regressão linear múltipla duas ou mais variáveis independentes são usadas para a previsão. Obteve-se os seguintes valores:

**MAE**: 0.0851479345146642.

**RMSE**: 0.10351671330958.

Visto que os valores são extremamente similares não se pode concluir que é possível obter melhores resultados utilizando o modelo de regressão linear múltipla. Ambos os modelos têm desempenho muito próximo um do outro.

*Exercício 7:* Tendo em conta o conjunto de dados apresentado, pretende-se prever o atributo “vo2\_results” relativo ao resultado do teste de volume de oxigênio máximo, aplicando.

*a)* Regressão linear múltipla.

Primeiro foi analisado a matriz de correlação e como foi observado apenas o “gender”, “pro\_level”, “winter\_camp”, “altitude\_results”, “hr\_results” seriam as variáveis com maior correlação. Estas variáveis foram escolhidas para melhorar a precisão dos resultados obtidos. Neste caso, para saber se realmente seriam uma influência estatisticamente significativa no “vo2\_results”.

Figura 9- Resultados da Regressão Linear Múltipla

Uma imagem com texto, captura de ecrã, Tipo de letra, menu

Descrição gerada automaticamente

Conclui-se que as variáveis “winter\_camp”, “altitude\_results” e “hr\_results” têm uma influência significativa no resultado “vo2\_results”, pois o seu “p-value” ser menor que o nível de significância “α=0.05”. O “gender” e “pro\_level” também mostram uma possível associação, mas com uma evidencia estatística superior ao “nível de significância”, portanto, desta forma não foram consideradas.

Equação da Reta obtida: “vo2\_results = -0.008440 + (-0.009374 \* gender) + (-0.007760 \* pro\_level) + (0.025133 \* winter\_camp) + (0.130902 \* altitude\_results) + (0.862144 \* hr\_results)”.

Figura 10 - Análise de Resíduos da Regressão Linear Múltipla

Uma imagem com texto, diagrama, file, Paralelo

Descrição gerada automaticamente

Pela “Figura 10” foi observado como cada variável influencia o valor de “v02\_results”. O “winter\_camp”, “altitude\_results” e o “hr\_results” tem um declive positivo e mais afastado de “0” e por isso o aumento nos valores destas variáveis estão associados a aumentos no “v02\_results”. Por outro lado, as variáveis “gender” e “pro\_level” têm um declive mais próximo de zero e negativo e isso sugere que essas variáveis têm uma influência relativamente fraca no resultado do “v02\_results”. Essas mudanças têm um efeito negativo e menos significativo no “vo2\_results”.

b) Árvore de regressão, usando a função rpart. Apresente arvore de regressão obtida.

Foi realizada a mesma abordagem como da alínea a) em termos de escolha de variáveis. Neste caso apenas a variável “hr\_results” foi selecionada como preditor. Desta maneira foi definido 3 dígitos para o “Plot”.

Figura 11 - Árvore de Regressão

Uma imagem com diagrama, file, Paralelo, texto

Descrição gerada automaticamente

Ao analisar o gráfico foi observado o relacionamento entre o preditor “hr\_results” e “vo2\_results”. Na arvore a única variável utilizada para fazer as divisões nos nós da arvore. Isto demonstra que no conjunto de dados utilizados, o valor de “hr\_results” é um forte indicador para prever o valor médio de “vo2\_results”.

À medida que o “hr\_results” aumenta espera-se que o valor médio de “vo2\_results” também aumente. Esta relação positiva entre “hr\_results” e “vo2\_results” indica uma possível associação entre a frequência cardíaca e o volume de oxigênio máximo.

c) Rede Neuronal usando a função neuralnet, fazendo variar os parâmetros. Apresente a rede obtida.

Como já tinha sido referido anteriormente a mesma abordagem foi efetuada para esta alínea. Assim foram utilizadas as mesmas variáveis para criar a rede neuronal.

Figura 12 - Rede Neuronal

Uma imagem com diagrama, file, texto, mapa

Descrição gerada automaticamente

*Exercício 8: Compare os resultados obtidos pelos modelos referidos na questão 7, usando o erro médio absoluto (MAE) e a raiz quadrada do erro médio (RMSE).*

Figura 13 - Resultados MAE & RMSE (Regressão Linear Múltipla)

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

No contexto da regressão linear múltipla o “MAE” indica que em média as previsões da regressão linear estão se a desviar “0.04177241” unidades do valor real de “vo2\_results”, já o “RMSE” o valor de “0.05171938” representa um desvio de aproximadamente “0.0517” unidades em relação aos valores reais de “vo2\_results”. Foi calculado também o “R^2” também conhecido como coeficiente de determinação e este demonstra que cerca de que aproximadamente “91,25%” da variabilidade dos dados de “vo2\_results” pode ser explicada pelas variáveis independentes incluídas neste modelo. Pode-se concluir que são resultados positivos por o “MAE” e o “RMSE” serem bastante baixos e “R^2” relativamente alto.

Figura 14 - Resultados MAE & RMSE (Árvore de Regressão)

Uma imagem com texto, Tipo de letra, captura de ecrã, tipografia

Descrição gerada automaticamente

No contexto da “Árvore de Regressão” o “MAE” indica que em média as previsões da regressão linear estão se a desviar “0.04798744” unidades do valor real de “vo2\_results”, já o “RMSE” o valor de “0.06048881” representa um desvio de aproximadamente “0.06048881” unidades em relação aos valores reais de “vo2\_results”. Ao calcular o R^2 obteve-se que era de “87.95%”. Comparativamente ao modelo anterior existiu uma piora nos resultados.

Figura 15 - Resultados MAE & RMSE (Rede Neuronal)

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

No contexto da “Rede Neuronal” o “MAE” indica que em média as previsões da regressão linear estão se a desviar “0.04211468” unidades do valor real de “vo2\_results”, já o “RMSE” o valor de “0.0520704” representa um desvio de aproximadamente “0.06048881” unidades em relação aos valores reais de “vo2\_results”. Ao calcular o R^2 obteve-se que era de “91.13%”. Comparativamente ao modelo da “Arvore” de Regressão” existiu uma melhoria nos resultados, mas com a “Regressão Linear Múltipla” obteve-se resultados piores.

*Exercício 9: Justifique se os resultados obtidos para os dois melhores modelos são estatisticamente significativos (para um nível de significância de 5%). Identifique o modelo que apresenta o melhor desempenho.*

Do ponto de vista de uma análise informal a “Regressão Linear Múltipla” apresenta resultados melhores que a “Rede Neuronal” em termos de “MAE”, “RMSE” e “R^2”. Foi realizado um teste formal chamado “t-test”. Foi tomado “H0” como os dois modelos fossem iguais e “H1” como se os modelos fossem diferentes. Tendo como resultado do “p-value” de “0.9007” e este sendo maior que o nível de significância não foi possível mostrar que existem diferenças significativas entre os dois modelos para se confirmar se existe um modelo melhor do que o outro.

*Exercício 4.2.1)* Estude a capacidade preditiva relativamente ao atributo “**Pro\_level**”.

Primeiramente, fez-se a árvore de decisão com os seguintes dados: altitude\_results, hr\_results, vo2\_results e gender.

Figura 16 - Árvore de decisão

Uma imagem com texto, captura de ecrã, diagrama

Descrição gerada automaticamente

De seguida, criou-se a rede neuronal, utilizando-se os mesmos dados da árvore de decisão. Obtiveram-se os seguintes resultados:

Figura 17 - Rede Neuronal

Uma imagem com diagrama, file, texto, mapa

Descrição gerada automaticamente

Finalmente, utilizou-se o modelo k-vizinhos-mais-próximos.

a) Usou-se o método k-fold cross validation na rede neuronal e obteve-se o seguinte resultado: média=59.5465193480903 e um desvio padrão=0.0267739621065555. Para o método k-vizinhos-mais-próximos obteve-se: média=60.776662950576 e um desvio padrão = 0.054814074102143.

b) O k-vizinhos-mais-próximos também é conhecido por algoritmo "Lazy Learning" porque o processamento dos exemplos de treino é adiado até a realização novamente de previsões, o treino consiste em armazenar apenas os dados de treino.

As implicações deste tipo de modelo são as seguintes: não requer um treino explícito, é sensível aos dados de treino e pode ser computacionalmente intensivo durante a fase de teste. É necessária uma escolha adequada do valor k para obter bons resultados.

c) Para este exercício foi feito um t.test onde:

H0: Não existe diferença significativa no desempenho dos dois modelos.

H1: Existe diferença significativa no desempenho dos dois modelos.

Para um nível de significância de 5%, com o pValue=0.5341 não temos evidências estatísticas para afirmar que há diferença significativa no desempenho dos dois modelos.

d) *Accuracy*:

* Árvore de decisão: 65%
* Rede neuronal: 69%
* K-vizinhos-mais-próximos: 67.33%

*Sensitivity*:

* Árvore de decisão: 79.17%
* Rede neuronal: 46.88%
* K-vizinhos-mais-próximos: 36.45%

*Specificity*:

* Árvore de decisão: 63.77%
* Rede neuronal: 71.64%
* K-vizinhos-mais-próximos: 92.23%

*F1*:

* Árvore de decisão: 26.5786%
* Rede neuronal: 24.3871%
* K-vizinhos-mais-próximos: 40.24%

Analisando os dados obtidos, conclui-se que para ***accuracy***a rede neuronal tem o melhor desempenho, para **sensitivity** é a árvore de decisão e para a **specificity** e o **F1** o K-vizinhos-mais-próximos.

*Exercício 4.2.2)* Estude a capacidade preditiva relativamente ao atributo “**Winter\_training\_camp**” utilizando árvore de decisão e rede neuronal.

a) Usou-se o método k-fold cross validation na árvore de decisão e obteve-se o seguinte resultado: média= 69.3772893772894 e um desvio padrão= 0.0430219096456042. Para a rede neuronal obteve-se: média=60.8322466891117 e um desvio padrão= 0.0238806625079828.

b) Para este exercício foi feito um t.test onde:

H0: Não existe diferença significativa no desempenho dos dois modelos.

H1: Existe diferença significativa no desempenho dos dois modelos.

Para um nível de significância de 5%, com o pValue=0.4617 não temos evidências estatísticas para afirmar que há diferença significativa no desempenho dos dois modelos.

c) *Accuracy:*

* Árvore de decisão: 67.67%
* Rede neuronal: 67.33%

*Sensitivity:*

* Árvore de decisão: 55%
* Rede neuronal: 52.73%

*Specificity:*

* Árvore de decisão: 69.62%
* Rede neuronal: 70.61%

*F1:*

* Árvore de decisão: 31.2034%
* Rede neuronal: 37.1778%

Analisando os resultados obtidos, conclui-se que os dois modelos obtêm performances bastantes parecidas. O único parâmetro onde se encontra uma diferença percentual acima de 5% é no F1 onde a rede neuronal é superior.

##### Conclusão

De acordo com os resultados obtidos não se consegue destacar um método que tenha um desempenho bastante superior aos restantes pois os resultados são muito parecidos e os valores variam constantemente.

O uso do R na análise de dados apresenta diversas vantagens, tais como a flexibilidade, disponibilidade de uma vasta variedade de ferramentas, pacotes e bibliotecas para diferentes tipos de análise. Facilita a visualização dos resultados através de gráficos personalizáveis e relatórios.

Durante a realização do trabalho em grupo, conseguirmos melhorar as nossas habilidades em R e técnicas aprendidas ao longo das aulas de Análise de Dados. Além disso, tivemos a oportunidade de expandir o nosso conhecimento na área, consolidando assim a aprendizagem.

##### References

(2021). Obtido de javatpoint.

Darmiton. (22 de 06 de 2020). *k-vizinhos mais próximos: uma análise.* Obtido de MAp: https://maquinasqueaprendem.com/2020/06/22/k-vizinhos-mais-proximos-uma-analise/

ISEP. (16 de Fevereiro de 2023). *Análise Exploratória de Dados - Estatística DescritivaFicheiro*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277049/mod\_resource/content/9/ANADI2122\_AED\_T.pdf

ISEP. (16 de Fevereiro de 2023). *Regressão Linear*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277082/mod\_resource/content/6/ANADI-LEI-1718\_RegressaoLinear.pdf

ISEP. (16 de Fevereiro de 2023). *Testes Paramétricos*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277061/mod\_resource/content/4/TestesParametricosPT.pdf

*O que é um diagrama de árvore de decisão?* (s.d.). Obtido de Lucidchart.

*Redes Neurais*. (s.d.). Obtido de SAS: https://www.sas.com/pt\_br/insights/analytics/neural-networks.html

Torgo, L. (03 de 10 de 2003). *Árvores de Regressão.* Obtido de Universidade do Porto: https://www.dcc.fc.up.pt/~ltorgo/SebentaR/HTML/node26.html