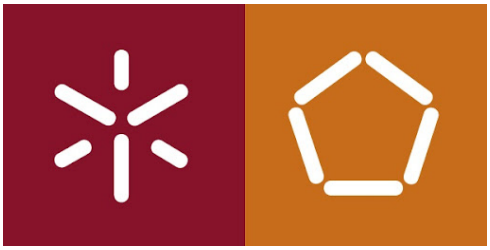

Universidade do Minho

Mestrado em Engenharia Informática

# Conceção e otimização de um modelo de Aprendizagem Profunda

Aprendizagem Profunda

Grupo 09

Gonçalo Almeida (pg47212)
Leonardo Marreiros (pg47398)
Maria Sofia Marques (pg47489)
Pedro Fernandes (pg47559)

19 de maio de 2022

# Conteúdo

**Lista de Figuras** **i**

**Lista de Tabelas** **ii**

**1 Previsão de idade cerebral** **1**

1.1 Domínio . . . 1
1.2 Objetivos . . . 1
1.3 Metodologia . . . 1
1.4 Análise dos Dados . . . 2
1.4.1 *Idade* . . . 2
1.4.2 Sexo . . . 2
1.4.3 Educação . . . 3
1.5 Pré-Processamento dos Dados . . . 3
1.6 Modelos *Deep Learning* . . . 4
1.6.1 Modelo 1 . . . 4
1.6.2 Modelo 2 . . . 6
1.6.3 Modelo 3 . . . 6
1.7 *Feature Importance* . . . 7
1.8 Interpretação dos resultados . . . 8
1.8.1 Modelo 1 . . . 8
1.8.2 Modelo 3 . . . 9
1.8.3 Comparação . . . 10
1.9 Considerações Finais . . . 11

# Lista de Figuras

1 *Hisplot* idades . . . 2
2 *Countplot* idades . . . 2
3 *Hisplot* sexo . . . 2
4 *Countplot* sexo . . . 2
5 *Boxplot* idade por sexo . . . 3
6 *Hisplot* educação . . . 3
7 *Countplot* educação . . . 3
8 *Boxplot* idade por educação . . . 3

11 Gráfico 2 . . . . . 5
9 Gráfico 1 . . . . . 5
10 Gráfico 2 . . . . . 5
12 Exemplo SHAP 1 . . . . . 8
13 Exemplo SHAP 2 . . . . . 8
14 Visualização de previsões do modelo 3 (Secção 1.6.3) . . . . . 8
15 Gráfico de erros residuais . . . . . 9
16 Gráfico de relação entre valor real e previsto . . . . . 9
17 MAE por grupo de idades do modelo 1 (Secção 1.6.3) . . . . . 9
18 Gráfico de erros residuais . . . . . 10
19 Gráfico de relação entre valor real e previsto . . . . . 10
20 MAE por grupo de idades do modelo 1 (Secção 1.6.3) . . . . . 10
21 Gráfico de erros residuais entre modelos . . . . . 11
22 Gráfico de comparação do MAE por grupo de idade . . . . . 11

# Lista de Tabelas

1 Tabela de comparação . . . . . 11

# 1 Previsão de idade cerebral

## 1.1 Domínio

O envelhecimento cerebral é um processo contínuo e que ocorre ao longo da vida. No entanto, as taxas individuais de envelhecimento são moldadas por uma grande variedade de interações entre fatores ambientais, genéticos e epigenéticos. A neurociência tem vindo a fornecer avaliações de risco e previsões para doenças neurodegenerativas e neuropsiquiátricas associadas à idade de um sujeito através do estabelecimento de biomarcadores. Para isso tem sido desenvolvidos estudos, os quais demonstraram que os sistemas de *machine learning* e *deep learning* podem ser muito úteis para prever com precisão a idade do cérebro ou a "lacuna cerebral", isto é, a diferença entre a idade cerebral prevista e a real, e desta forma ajudar no estabelecimento de disgnósticos precoces de doenças neurodegenerativas.

## 1.2 Objetivos

O objetivo principal do desenvolvimento deste trabalho é desenvolver e otimizar um modelo de aprendizagem profunda capaz de prever a idade do cérebro a partir de características de conectividade estrutural a partir da RM de difusão.

Em conformidade com aquilo que já foi dito, procuramos prever a "lacuna cerebral"pelo que a métrica com a qual estaremos mais preocupados será o *Mean Squared Error* visto retratar exatamente este conceito, isto é, a diferença quadrática média entre os valores estimados e o valor real. Além disso, pretende-se analisar os resultados obtidos e identificar eventuais sujeitos para os quais o modelo tem dificuldade em fazer a previsão assim como apurar quais as conexões cerebrais mais relevantes para a previsão da idade do cérebro.

## 1.3 Metodologia

A metodologia de extração de conhecimento utilizada foi a CRISP-DM. Começamos por reconhecer como fatores como o sexo e anos de educação afetam a possibilidade de envelhecimento retardado e os valores que seriam de esperar (idade cerebral menos avançada) para os sujeitos que de facto tiveram maior número de anos de educação por exemplo. De seguida, procuramos perceber os dados referentes à conectividade estrutural do cérebro e descobrimos que representam conexões entre zonas do cérebro pelo que, estando na forma de matriz, metade dos dados são redundantes uma vez que se existe conexão entre uma zona A e B então também haverá entre B e A. Ao visualizarmos os dados verificamos que havia uma maior quantidade de dados de sujeitos na faxa etária entre 13 a 27 anos e 54 a 79 anos e o facto de serem poucos dados o que poderia causar problemas de *overfitting*. Com isto, passamos para a fase de processamento dos dados onde inicialmente separamos a coluna das idades das colunas de educação e sexo que seriam usadas apenas numa próxima iteração. Passamos então para a modelação onde reparamos que uma vez que os dados eram poucos, talvez não fosse necessário várias camadas ou elevados valores

de *dropout* o geraria melhores resultados e voltamos a treinar os modelos. Ao analisar os resultados que obtemos procuramos adicionar ou remover camadas e hiperparâmetros até obter um modelo apropriado. Após esta fase, decidimos incorporar as *features* representantes do sexo e educação dos sujeitos. Finalmente, antes de tirar conclusões sobre os resultados, voltamos a rever o contexto do problema e confirmar que de facto os objetivos que tínhamos definidos foram alcançados.

## 1.4 Análise dos Dados

O *dataset* em questão é referente a matrizes de conectividade estrutural extraídas de *scans* de DTI-MRI e opcionalmente algumas informações sobre os sujeitos. A primeira análise a ser realizada foi visualizar os tipos de dados assim como verificar o seu balanceamento.

### 1.4.1 *Idade*

Esta coluna representada por um dado discreto, é referente à idade de cada pessoa presente no *dataset*. Os dados incluem 112 sujeitos com idades entre os 13 e os 79 anos, com um valor médio de 44.312500 ± 22.540848anos o que significa que há um bom balanceamento dos dados. Tal pode ser comprovado pelo valor de *skewness* que é -0.048943.

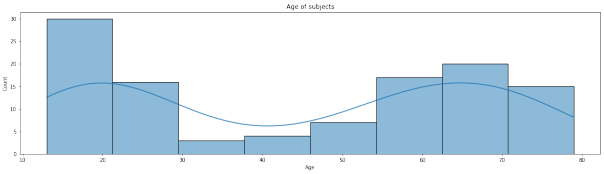


Figura 1: *Hisplot* idades

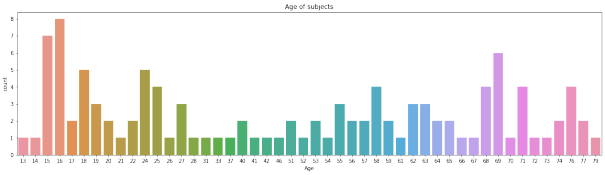


Figura 2: *Countplot* idades

### 1.4.2 Sexo

Este atributo binário representa o sexo do sujeito. Os dados incluem 112 sujeitos com sexo igual a 0 ou 1. O valor 0 representa sexo masculino, o valor 1 representa o sexo feminino. Mais uma vez, os dados apresentam um bom balanceamento com um valor de *skewness* igual a 0.071474. Além disso, com a criação de um *boxplot* podemos verificar que a média de idades dos sujeitos é bastante mais baixa para os sujeitos femininos.

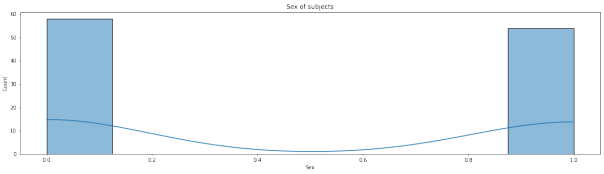


Figura 3: *Hisplot* sexo

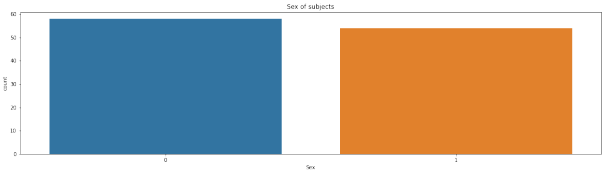


Figura 4: *Countplot* sexo

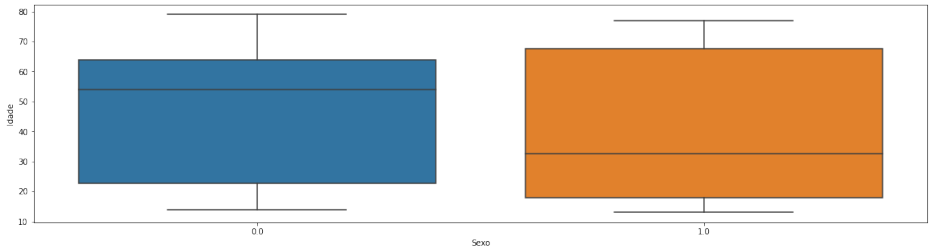


Figura 5: *Boxplot* idade por sexo

### 1.4.3 Educação

Este atributo discreto representa o número de anos de educação do sujeito. Os dados incluem 112 sujeitos com educação entre 0 e 20 anos, com um valor médio de 9.035714 ± 4.887902anos. O valor do *skewness* é igual a 0.292163 pelo que há uma ligeira tendência à esquerda. Além disso, com a criação de um *boxplot* podemos verificar que de forma geral uma menor educação está associada a maior idade.

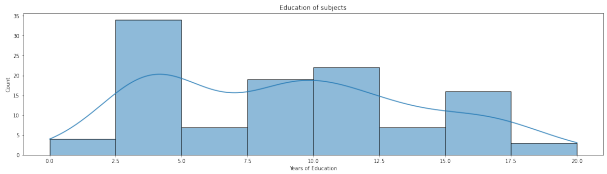


Figura 6: *Hisplot* educação

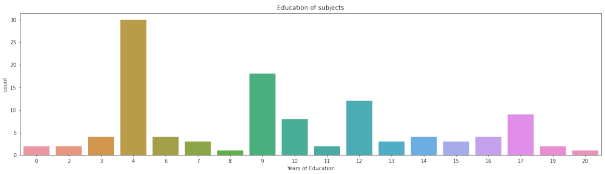


Figura 7: *Countplot* educação

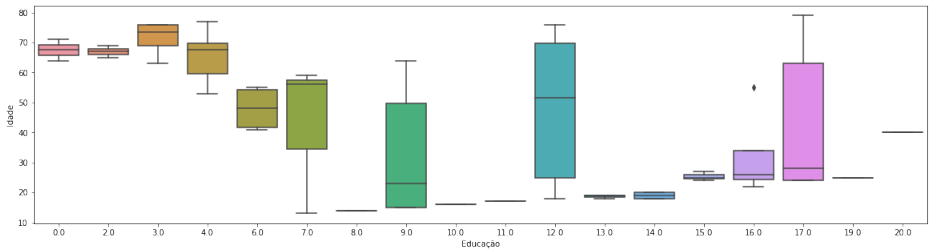


Figura 8: *Boxplot* idade por educação

## 1.5 Pré-Processamento dos Dados

Uma vez que os dados das matrizes de conexão já se encontravam limiarizados e normalizados, não foi feito qualquer tratamento. Quanto aos restantes dados, numa das iterações, os dados de educação foram normalizados, contudo, como esta implementação não pareceu melhorar os resultados ou o tempo de convergência, acabamos por optar por não o fazer.

Sendo assim, o único tratamento que foi efetuado foi um *shuffle* aos dados que será justificado na secção 1.6.1.

## 1.6 Modelos *Deep Learning*

### 1.6.1 Modelo 1

Numa primeira iteração, não foram utilizados os dados de sexo e anos de educação treinando um modelo apenas com os dados de conetividade estrutural do cérebro. Passamos por várias etapas de *tuning* do modelo, acrescentando e retirando camadas e ajustando parâmetros. No final, acabamos com um modelo com poucas camadas e baixo número de filtros para prevenir o *overfit* mas com apenas uma camada de *dropout* já que os dados eram poucos.

```
def build_model(num_classes, activation='relu', loss='mean_squared_error'):
    reset_random_seeds()
    model = tf.keras.Sequential()
    model.add(L.Conv2D(32,(3,3),input_shape=(90,90,1),padding='same',activation='relu'))
    model.add(L.MaxPool2D(2,2))
    model.add(L.BatchNormalization())

    model.add(L.Conv2D(64,(3,3),activation='relu'))
    model.add(L.MaxPool2D(2,2))

    model.add(L.Conv2D(128,(3,3),activation='relu'))
    model.add(L.MaxPool2D(2,2))

    model.add(L.Flatten())

    model.add(L.Dense(64, activation='relu')),
    model.add(L.Dropout(0.2)),
    model.add(L.Dense(num_classes,activation="relu"))

    sgd = tf.keras.optimizers.SGD(momentum=0.9)

    model.compile(
        optimizer='adam',
        loss=loss,
        metrics=['mae']
    )

    return model
```

Numa primeira fase, o modelo estava a ser treinado com um $validation\_split = 0.1$ pelo que os últimos 10% dos dados estavam a ser usados para validação. Ao analisar os resultados deste treino, foi possível verificar que o erro para os casos de treino era muito baixo em comparação com o erro nos casos de validação, um caso claro de *overfitting*. Com isto, decidimos usar a totalidade dos dados para treino, uma vez que estes já eram reduzidos e ao usar um set de validação perdiamos dados essenciais.

Uma vez que agora não tínhamos qualquer validação da performance do modelo treinado, decidimos utilizar *KFold* permitindo desta forma verificar o nosso modelo com diferentes dados. Antes de aplicar esta técnica, foi feito um *shuffle* aos dados uma vez que estes estavam organizados de grosso modo por ordem crescente de idades, o que significa que as *folds* criadas para validar o modelo contêm dados mais representativos da realidade.

```
------------------------------------------------------------------------
Score per fold
------------------------------------------------------------------------
> Fold 1 - Loss: 105.65406036376953 - MAE: 8.584004402160645
------------------------------------------------------------------------
> Fold 2 - Loss: 109.27930450439453 - MAE: 7.8008952140808105
------------------------------------------------------------------------
> Fold 3 - Loss: 108.26921081542969 - MAE: 7.761288642883301
------------------------------------------------------------------------
> Fold 4 - Loss: 107.8408432006836 - MAE: 8.097200393676758
------------------------------------------------------------------------
> Fold 5 - Loss: 174.97239685058594 - MAE: 10.137221336364746
------------------------------------------------------------------------
> Fold 6 - Loss: 104.7122573852539 - MAE: 7.785461902618408
------------------------------------------------------------------------
> Fold 7 - Loss: 106.06411743164062 - MAE: 8.03720474243164
------------------------------------------------------------------------
> Fold 8 - Loss: 94.9795150756836 - MAE: 7.327514171600342
------------------------------------------------------------------------
> Fold 9 - Loss: 92.38077545166016 - MAE: 8.465104103088379
------------------------------------------------------------------------
> Fold 10 - Loss: 96.297119140625 - MAE: 7.6707563400268555
------------------------------------------------------------------------
Average scores for all folds:
> MAE: 8.16666512489319 (+- 0.7457511971127161)
> Loss: 110.04496002197266
> Epochs: 371.7
------------------------------------------------------------------------
```

Figura 11: Gráfico 2

Para cada fold é treinado o modelo e gerado um gráfico para avaliar o *overfitting* (Figuras 21, 22). Foi criado um *callback* personalizado que é utilizado para prevenir o *overfit* e guardado o número de *epochs* a que o treino chegou. No final do treino das *folds* é gerado um sumário dos resultados de cada *fold* e uma média destes resultados (Figura 20). O resultado da média de *epochs* é depois utilizado quando é treinado o modelo com a totalidade dos dados.

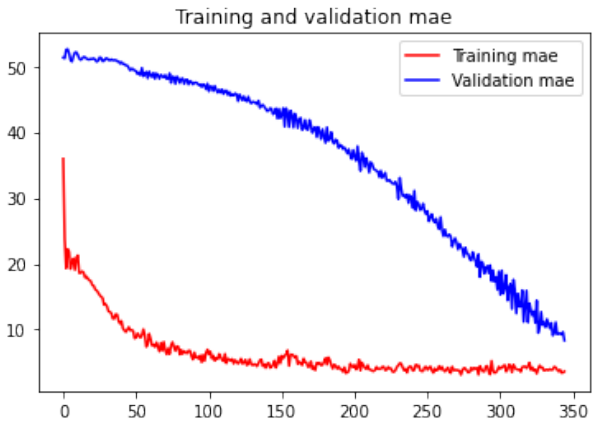


Figura 9: Gráfico 1

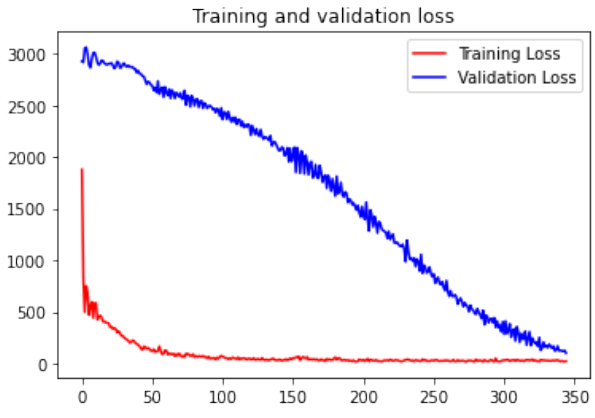


Figura 10: Gráfico 2

### 1.6.2 Modelo 2

Numa segunda iteração, tínhamos como objetivo ter em consideração as *features* de sexo e educação nas nossas previsões. Para isto construímos um novo modelo que recebia como *input* o sexo, a educação e a idade prevista pelo primeiro modelo. O modelo construído foi o seguinte:

```
def build_modelMore(num_classes, activation='relu', loss='mean_squared_error'):
    model = tf.keras.Sequential()

    model.add(L.Dense(64, input_dim=3, activation='relu'))

    model.add(L.Dense(32, activation='relu'))

    model.add(L.Dense(8, activation='relu'))
    model.add(L.Dropout(0.2))
    model.add(L.Dense(1, activation='relu'))

    sgd = tf.keras.optimizers.SGD(momentum=0.9)

    model.compile(
        optimizer='adam',
        loss=loss,
        metrics=['mae']
    )

    return model
```

Desta forma, para obter as previsões finais era necessário passar por um *pipeline*: treinar o modelo da secção anterior e obter as previsões deste para os dados de teste; a este resultado eram concatenados os valores correspondentes de sexo e anos de educação; o novo modelo era então treinado com este novo conjunto de dados.

Ao analisar o resultado desta solução reparamos a qualidade das previsões era pior do que com a utilização de apenas um modelo. Rapidamente chegamos à conclusão que esta não era uma boa solução uma vez que os resultados insatisfatórios advinham não só de desvios do valor real do primeiro modelo como a propagação e acentuação desses mesmos desvios no segundo modelo.

### 1.6.3 Modelo 3

Finalmente, numa última iteração, desenvolvemos um modelo que toma em consideração não só as matrizes de conexões como também os dados de sexo e educação. O modelo é o seguinte:

```
def build_model2(num_classes, activation='relu', loss='mean_squared_error'):
    reset_random_seeds()

    first_input= L.Input((90,90,1))
```

```
second_input = L.Input((2,))
model = L.Conv2D(32,(3,3),padding='same',activation='relu')(first_input)
model = L.MaxPool2D(2,2)(model)
model = L.BatchNormalization()(model)

model = L.Conv2D(64,(3,3),activation='relu')(model)
model = L.MaxPool2D(2,2)(model)

model = L.Conv2D(128,(3,3),activation='relu')(model)
model = L.MaxPool2D(2,2)(model)


model = L.Flatten()(model)
model = L.concatenate([model,second_input], axis=1)

model = L.Dense(64, activation='relu')(model)
model = L.Dropout(0.2)(model)
predictions = L.Dense(num_classes,activation="relu")(model)


model = Model(inputs=[first_input,second_input], outputs=predictions)

model.compile(
    optimizer='adam',
    loss=loss,
    metrics=['mae']
)

return model
```

Após a layer Flatten, foi inserida uma nova camada, *concatenate*, que recebe o vetor com as *features* das matrizes e lhes junta um novo vetor label com as novas características.

## 1.7 *Feature Importance*

De forma a apurar quais foram as conexões cerebrais mais relevantes para a previsão da idade do cérebro foi implementado o SHAP *(SHapley Additive exPlanations)*. SHAP é uma abordagem da teoria dos jogos para explicar o *output* de qualquer modelo de *machine learning*.

Neste caso foi utilizado um *DeepExplainer* que é um algoritmo de aproximação de alta velocidade para valores SHAP em modelos de aprendizagem profunda que se baseia numa conexão com o DeepLIFT.

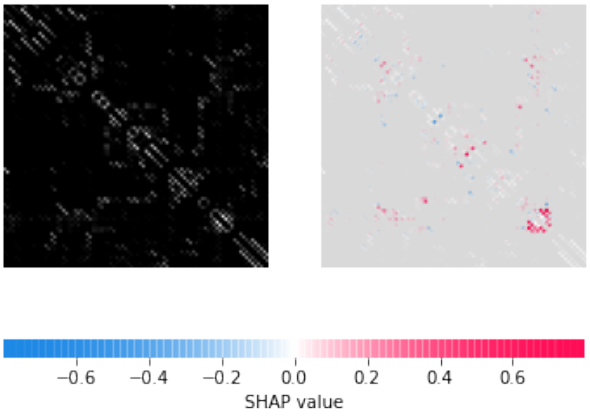


Figura 12: Exemplo SHAP 1

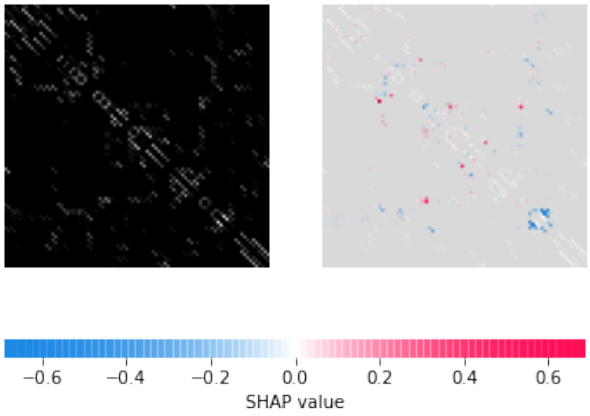


Figura 13: Exemplo SHAP 2

Os exemplos acima explicam os *outputs* para duas matrizes de dados referentes a sujeitos com idades à volta dos 70 e 25 anos respetivamente. Os pixels vermelhos aumentam o *output* do modelo enquanto os pixels azuis diminuem o *output*. É possível reparar que para diferentes idades as mesmas conexões podem contribuir positivamente ou negativamente para o *output* o que não é surpreendente pois os padrões encontrados num cérebro mais jovem com certeza não são os mesmos encontrados em cérebros mais idosos.

## 1.8 Interpretação dos resultados

Para cada um dos modelos 1 e 3 (Secções 1.6.1 e 1.6.3) foram analisados os resultados e feita uma análise crítica. Por fim foi feita uma comparação entre os dois modelos. Os resultados obtidos foram feitos com base nas previsões para próprios dados de treino dado o facto de não termos acesso ao valor real das previsões para os dados de teste.

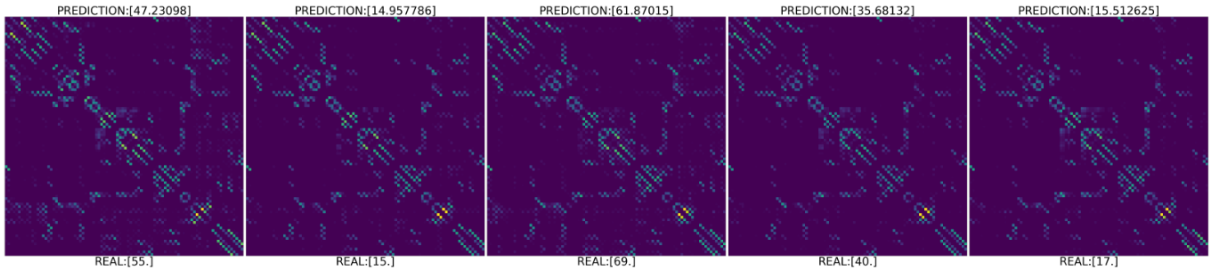


Figura 14: Visualização de previsões do modelo 3 (Secção 1.6.3)

### 1.8.1 Modelo 1

Para o modelo que não considera os dados de sexo e educação, foi feito um gráfico de erros residuais, isto é, a diferença entre o valor real e o valor previsto e um gráfico para ilustrar o valor previsto em relação com o valor real.

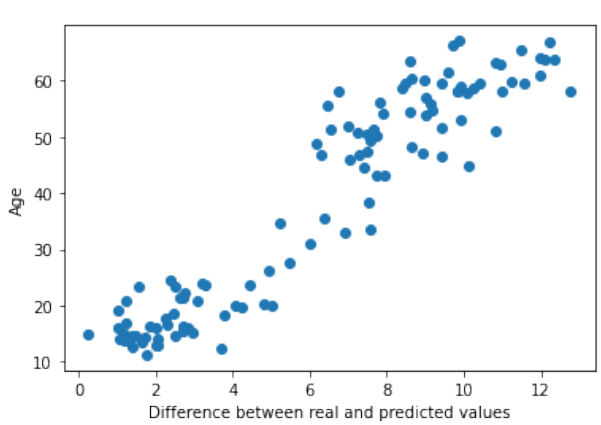


Figura 15: Gráfico de erros residuais

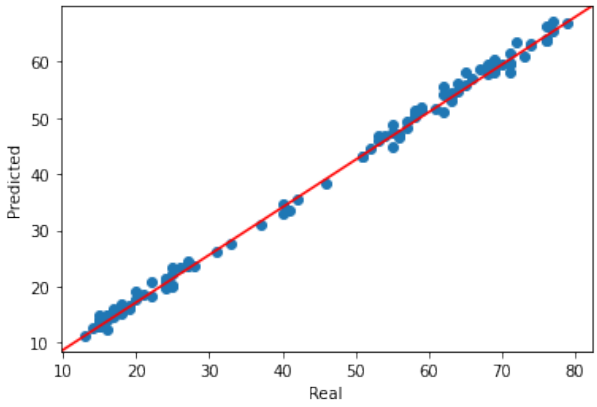


Figura 16: Gráfico de relação entre valor real e previsto

De modo geral é possível verificar que os resultados são satisfatórios, no entanto, também é possível verificar que à medida que as idades aumentam, a qualidade das previsões torna-se cada vez pior. Para comprovar isto ordenamos os dados por ordem crescente de idades e calculamos o MAE para cada grupo de idades (onde o grupo 1 representa o intervalo de 10-20, o grupo 2 o grupo 20-30,...) e criamos o seguinte gráfico que comprova o que foi especulado:

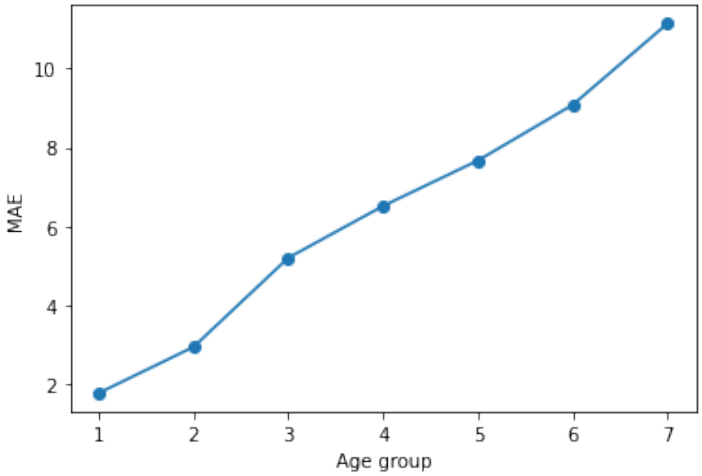


Figura 17: MAE por grupo de idades do modelo 1 (Secção 1.6.3)

### 1.8.2 Modelo 3

Igualmente, para o modelo que considera os dados de sexo e educação foi feito um gráfico de erros residuais e um gráfico para ilustrar o valor previsto em relação com o valor real.

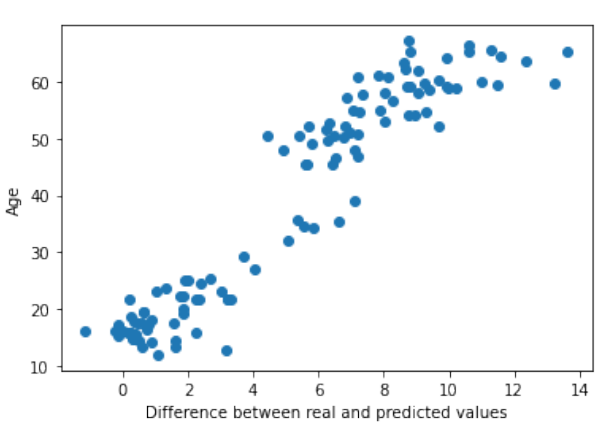


Figura 18: Gráfico de erros residuais

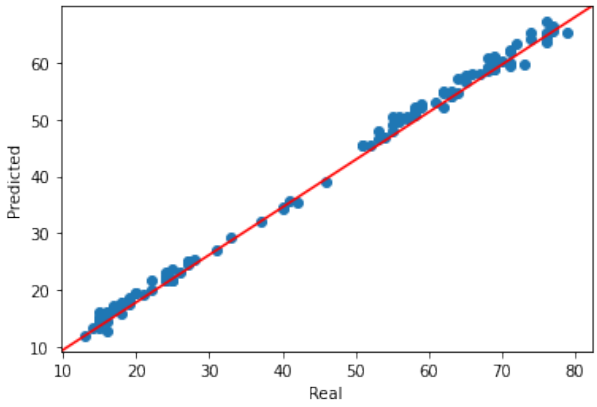


Figura 19: Gráfico de relação entre valor real e previsto

De modo geral é possível verificar que os resultados são idênticos ao modelo anterior o que seria de esperar pois, independentemente da importância que estes dados possam ter para a determinação da idade cerebral de um sujeito, o seu peso no modelo era muito baixo (2 features adicionais a uma matriz com 90x90=8100 valores). Mais uma vez, foi calculado o MAE por grupo de idades para ser possível comparar os dois modelos.

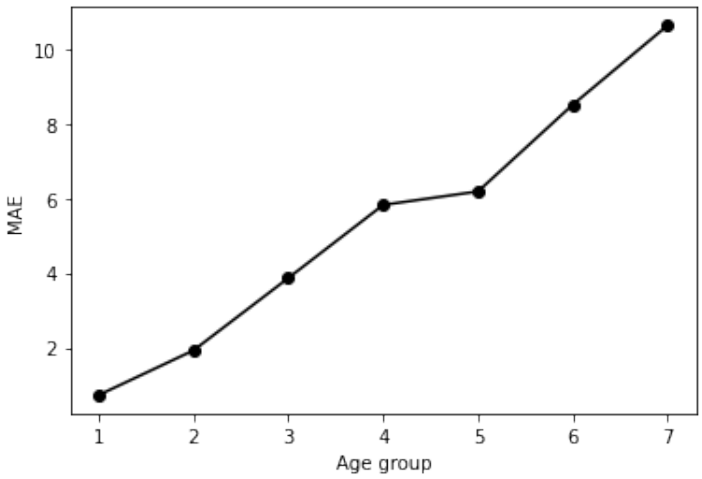


Figura 20: MAE por grupo de idades do modelo 1 (Secção 1.6.3)

### 1.8.3 Comparação

Foram criados dois gráficos para visualizar a performance dos dois modelos, um com a diferença entre as previsões de um modelo e do outro e o outro com ambas as retas de MAE por grupo de idade.

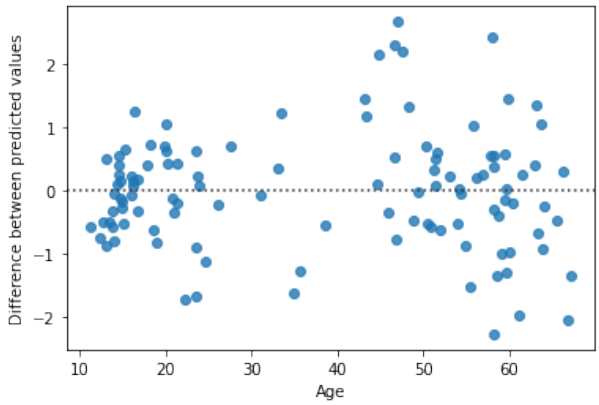


Figura 21: Gráfico de erros residuais entre modelos

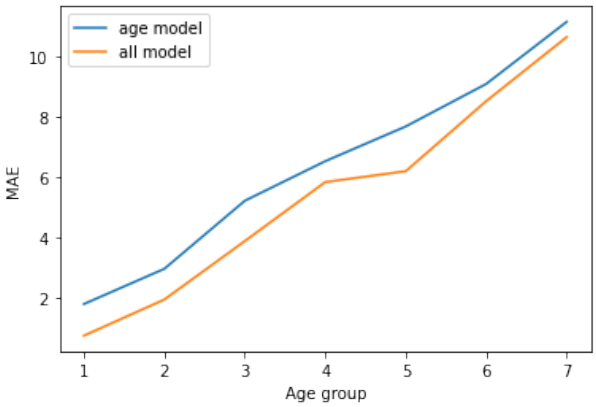


Figura 22: Gráfico de comparação do MAE por grupo de idade

De seguida apresenta-se uma instância dos resultados obtidos para cada um dos modelos:

| | **Modelo 1** | **Modelo 3** |
|---|---|---|
| **MAE global** | 6.1490 | 5.2113 |
| **MAE idades 10-20** | 1.7759 | 0.7270 |
| **MAE idades 20-30** | 2.9455 | 1.9258 |
| **MAE idades 30-40** | 5.2047 | 3.8662 |
| **MAE idades 40-50** | 6.5179 | 5.8249 |
| **MAE idades 50-60** | 7.6705 | 6.1859 |
| **MAE idades 60-70** | 9.0820 | 8.5126 |
| **MAE idades 70-80** | 11.144 | 10.638 |

Tabela 1: Tabela de comparação

Apesar de nesta instância o modelo 3 ter tido melhores resultados, este resultado não é constante. Dada a pequena dimensão dos dados, os resultados de cada modelo variam bastante entre instâncias. Para tentar mitigar este fator de aleatoriedade os modelos foram corridos várias vezes e, como previsto, os resultados variaram bastante: tanto o modelo 1 tinha melhores resultados como o modelo 3, mas a diferença absoluta entre eles nunca variava muito.

Posto isto, não é possível concluir se a idade e o sexo apresentam um papel fundamental na determinação da idade cerebral do cérebro de um sujeito, contudo, é possível concluir que, seja a contribuição da adição destas duas *features* positiva ou negativa, o seu impacto é sempre pequeno.

## 1.9 Considerações Finais

Neste trabalho apresentamos dois modelos de previsão da idade cerebral: um deles baseado apenas em matrizes de conetividade estrutural do cérebro; o outro com consideração a

dados de anos de educação e sexo dos sujeitos. Estes modelos apresentam previsões satisfatórias, tendo resultados progressivamente piores para idades superiores, o que não é de surpreender. Isto acontece, possivelmente, pois enquanto que os cérebros jovens devem apresentar características semelhantes, facilitando o reconhecimento de padrões para prever idades, os cérebros de sujeitos com idades mais avançadas devem ser mais heterogéneos na medida em que já viveram mais anos, possivelmente já passaram por traumas ou depressões ou outras condições que possam afetar o cérebro ou até apresentam doenças neuro degenerativas como Alzheimer ou Parkinson, mais presentes nestas faixas etárias.

Quanto à influência de anos de educação e sexo, apesar de dada a pequena dimensão dos dados não permitir conclusões sólidas, infere-se que a sua influência é baixa na previsão da idade cerebral. De facto, um estudo recente [1] contesta a crença comum de que obter uma educação superior pode ajudar a retardar o envelhecimento cerebral. Segundo alguns artigos, efeitos do envelhecimento podem ser mais aparentes em homens do que mulheres, no entanto não conseguimos chegar a essa conclusão com o nosso estudo.

Existiam alguns indivíduos com doenças cerebrais no *dataset* fornecido, e para os tentar identificar tentámos ver quais os casos em que a diferença entre o valor obtido no treino e real eram mais distantes. Para isto verificamos para quais sujeitos era maior o erro residual. Com isto, descobrimos que em cerca de 95% dos testes realizados, os indivíduos dos índices 87 e 96 dos dados de treino com idades 79 e 71 respetivamente, se encontravam entre as quatro piores previsões pelo que é provável que façam parte deste conjunto de sujeitos.

Dito isto, consideramos que, apesar da dimensão deste conjunto de dados ser baixa, foi possível desenvolver um modelo de previsão bastante satisfatório bem como tirar outras conclusões acerca dos dados. Como trabalho futuro gostaríamos de remover a informação redundante das matrizes e remover os valores de conexões que nunca existem (são sempre 0), colocando tudo isso num vetor e posteriormente utilizando camadas convolucionais de uma dimensão no modelo e estudar os resultados destas alterações.

[1] https://www.pnas.org/doi/full/10.1073/pnas.2101644118