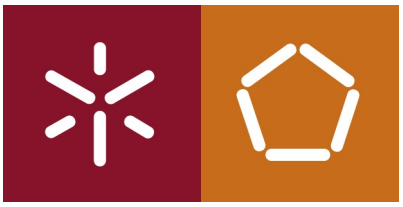

Universidade do Minho
mestrado integrado em engenharia informática

Computação Natural

Sistemas Inteligentes

(2º Semestre / 4º Ano)

Redes Neuronais Artificiais em Exames Clínicos

GRUPO 10
Diogo Braga (a82547)
Pedro Ferreira (a81135)
Ricardo Caçador (a81064)
Ricardo Veloso (a81919)

Abril, 2020

Conteúdo

1 Introdução

O presente relatório detalha a implementação de um sistema para determinar se tumores são ou não malignos. Este projeto foi proposto no âmbito da Unidade Curricular de Computação Natural, inserida no perfil de especialização de Sistemas Inteligentes.

Como é conhecimento geral, o cancro é uma das principais causas de morte no mundo inteiro, estando apenas atrás das doenças cardiovasculares, no que diz respeito à taxa de mortalidade. Esta tendência também se verifica no cenário Português, onde os tumores malignos representam cerca de 24.6% das mortes [1]. Por estas mesmas razões, estudos nesta área têm sido extremamente explorados, revelando-se um verdadeiro desafio para os investigadores.

Atualmente não existem mecanismos altamente especializados para a deteção de cancro. Ainda inferiores, são os mecanismos de deteção antecipada, que permite reduzir de uma forma significativa a necessidade de aplicar tratamentos de alta intensidade, o que acarreta melhorias para o estado físico e mental do doente oncológico.

Sendo assim, as ecografias são uma forma de detetar padrões anormais em tumores, levantando assim a suspeita de cancro. Ainda que as ecografias tenham uma eficácia relativamente alta, a maior parte das vezes é necessário recorrer a biópsias para confirmar se de facto o tumor é ou não maligno.

Dados os avanços registados na ao longo das últimas décadas na Intilêngia Artificial, com especial destaque para o *Machine Learning*, têm surgido aplicações desta área ao problema enunciado. Assim, o *Machine Learning* tem sido extremamente utilizado para a realização de diagnósticos de cancro e já se provou até que, com a ajuda deste, a taxa de acerto aumenta significativamente.

Neste trabalho, detalha-se a criação de uma Rede Neuronal Artificial (RNA) que permita prever, com eficácia, a natureza de um tumor. Esta eficácia deverá ser ajustada com o auxílio de Algoritmos Genéticos (AG), que irão permitir a otimização da RNA criada pelo grupo.

Em suma, o sistema terá como objetivo classificar um conjunto de dados como tumor maligno ou benigno, tendo por base uma RNA otimizada através da implementação de AG.

2 Análise, compreensão e tratamento inicial dos dados

Para que seja possível fazer uma boa exploração e um bom tratamento dos dados, é essencial compreendê-los. Para isso é necessário compreender o significado de cada *feature* e o que cada valor destas representa. Consequentemente, vamos apresentar uma tabela que incluí todas as *features*, a sua descrição, os seus respetivos valores e a interpretação destes.

FEATURE	**DESCRIÇÃO**	**VALOR**	**CORRESPONDÊNCIA**
BI-RADS	Classificação desenvolvida para ser utilizada originalmente com a mamografia	1	"Negativo"
		2	"Benigno"
		3	"Provavelmente benigno"
		4	"Suspeita de anormalidade"
		5	"Sugestivo de malignidade"
AGE	Idade do Paciente (em anos)		
SHAPE	Forma do Tumor	1	"Redonda"
		2	"Oval"
		3	"Lobular"
		4	"Irregular"
MARGIN	Margem do Tumor	1	"Circunscrita"
		2	"Microlobulada"
		3	"Obscurecida"
		4	"Mal definida"
		5	"Espiculada"
DENSITY	Densidade do Tumor	1	"Alta"
		2	"Constante"
		3	"Baixa"
		4	"Contém Gordura"
SEVERITY	Classificação do Tumor	0	"Benigno"
		1	"Maligno"

Após esta análise inicial foi automaticamente depreendido que a *feature* **BI-RADS** não é relevante para a classificação dos dados já que esta não é de natureza classificativa. Consequentemente, foi removida do *dataset*.

De seguida deparámo-nos com a existência de valores nulos representados por "?". Convertemos então os valores nulos por "NaN".

	BI_RADS	Age	Shape	Margin	Density	Severity
0	5	67	3	5	3	1
1	4	43	1	1	?	1

Figura 1: Antes das alterações

1	4	43	1	1	NaN	1

Figura 2: Após alterações

3 Análise e tratamento de valores nulos

Após a alteração na representação dos valores nulos (figura 2), passamos a uma parte importantíssima no que diz respeito à preparação dos dados: a análise e tratamento de valores nulos.

3.1 Análise dos valores nulos

Para realizar uma análise extremamente pormenorizada, optamos por observar os valores nulos através de diferentes perspectivas:

- **Valores nulos por *feature***:

Age	5
Shape	31
Margin	48
Density	76

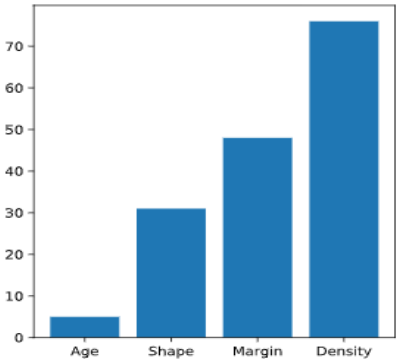

Figura 3: Valores absolutos e gráfico de barras de valores nulos por *feature*

Aqui podemos observar que a *feature* com o maior número de valores nulos é a **Density** com 76, e que a *feature* com menor número de valores nulos é a **Age**, com apenas 5.

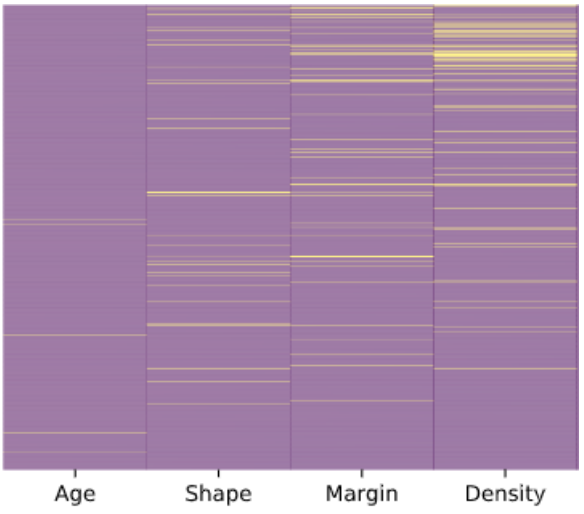

Figura 4: *Heatmap* de valores nulos por *feature*

Podemos confirmar esta observação analisando o *heatmap* da figura acima. É percetível que a densidade de valores nulos é abundante na *feature* **Density** e escassa na *feature* **Age**.

percent_missing	
Age	0.520833
Shape	3.229167
Margin	5.000000
Density	7.916667

Figura 5: Percentagens de valores nulos por *feature*

Para efeitos de confirmação, ainda verificamos as percentagens de valores nulos por *feature*. Como era de esperar, tudo se mantém.

- **Valores nulos por linha**:

	data_missing
613	2

Figura 6: Número de valores nulos por linha

Aqui é analisado o número de valores nulos por linha. Isto permite-nos saber quais as linhas com, por exemplo, 2 NaN. Esse é exatamente o exemplo apresentado em cima em que se verifica que a linha 613 tem 2 NaN.

- **Valores nulos por classe (Severity)**:

Severity	Age	Shape	Margin	Density
0.0	0	19	37	54
1.0	5	12	11	22

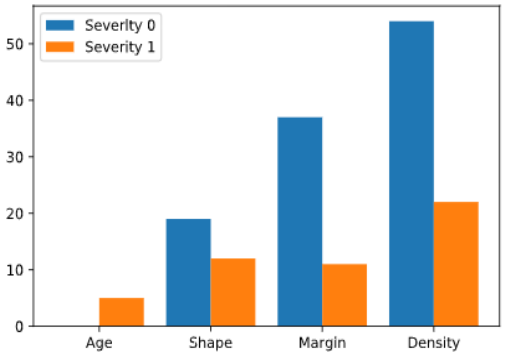

Figura 7: Valores absolutos e gráfico de barras de valores nulos por classe

Podemos verificar que existe um maior número de missing values quando a classificação é 0. A *feature* **Density** tem o maior número de valores nulos em qualquer classe.

- **Número de linhas com (x) NaN por classe**:

	1 NaN	2 NaN
Severity 0	66	22
Severity 1	34	8

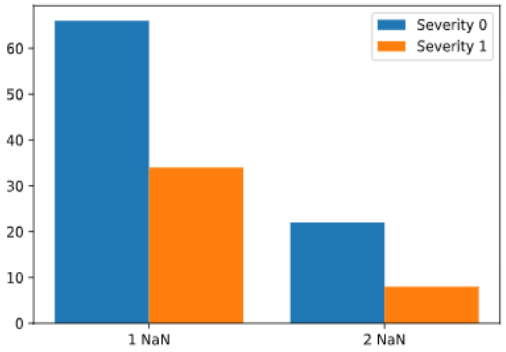

Figura 8: Valores absolutos e gráfico de barras do número de linhas NaN por classe

Podemos verificar que a maior parte das linhas com valores nulos têm apenas 1 NaN. Para uma classificação de 0, os valores continuam a ser extremamente superiores do que para uma classificação de 1 tanto para 1 NaN como para 2 NaN.

	Sum
Severity 0	88
Severity 1	42

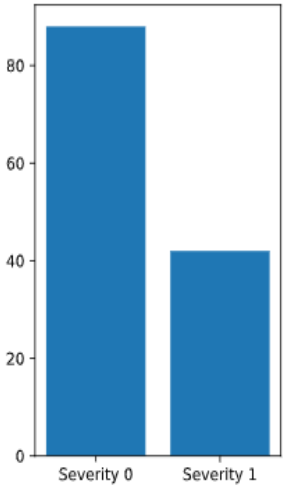

Figura 9: Valores absolutos e gráfico de barras do número de linhas NaN por classe

Tal como foi já deduzido anteriormente, confirma-se a existência de um maior número de linhas com valores nulos para a classe 0 do que para a classe 1.

3.2 Tratamento dos valores nulos

Após a realização desta análise exaustiva, o grupo observou que os valores nulos não são tendenciosos. Isto significa que os valores nulos parecem estar distribuídos aleatoriamente.

Esta conclusão pode parecer ligeiramente inconsistente com o que está referido em cima mas o que o grupo deduziu é que, ainda que exista uma disparidade entre a classe 0 e a classe 1, este valor deve-se à existência de um maior número de amostras para a classe 0 (516) do que para a classe 1 (445).

Sendo assim, o grupo decidiu seguir uma abordagem que nos garantisse a permanência de algumas destas linhas com valores NaN mas que não enviesasse os dados.

- Substituir os NaN das linhas com apenas 1 valor nulo

  Aqui optamos por calcular a moda dos valores para cada classe e depois substituir os NaN tendo em conta a sua classe e a moda associada a esta.

```
   Age  Shape  Margin  Density  Severity
0  46.0    1.0     1.0      3.0       0.0
   Age  Shape  Margin  Density  Severity
0  67.0    4.0     4.0      3.0       1.0
```

Figura 10: Moda de cada *feature* por classe

- Apagar linhas com 2 NaN

  Através da utilização da função `dropna()`.

4 Análise exploratória dos dados

A Análise Exploratória de Dados (EDA) é um processo de análise de dados que tem como objetivo descobrir as informações ocultas no conjunto de dados usando ferramentas estatísticas, álgebra linear e outras técnicas. Ajuda a entender melhor os dados e a destacar as suas principais características que podem ajudar a fazer previsões.

Começamos então por avaliar o equilíbrio do *dataset*.

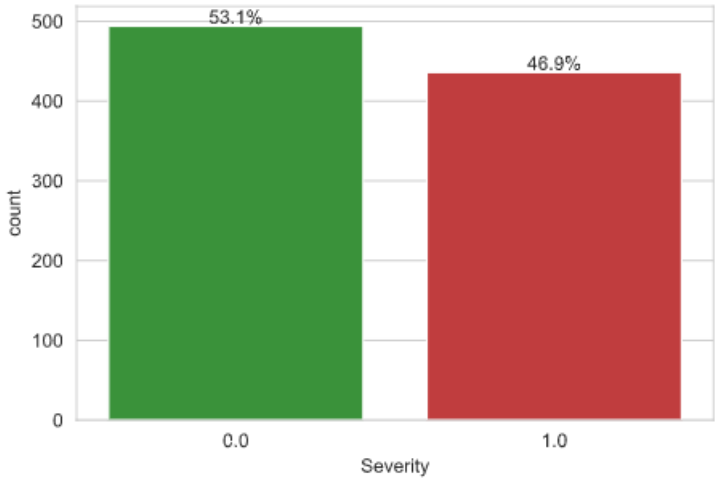

Figura 11: Equilíbrio do *dataset*

Observando este gráfico, podemos verificar que o dataset é de facto bastante equilibrado, apresentando um total de 53.1% de amostras para tumores benignos e 46.9% para tumores malignos.

De seguida, analisámos a idade no geral e por classificação dos tumores.

mean	55.641935
std	14.527127
min	18.000000
25%	45.000000
50%	57.000000
75%	66.000000
max	96.000000

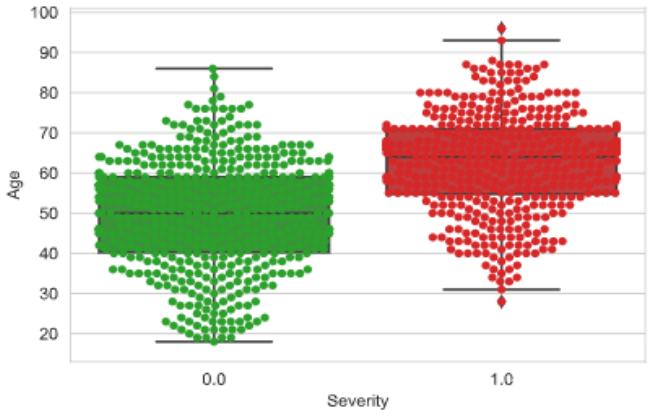

Figura 12: Métricas e *swarm-plot* das idades

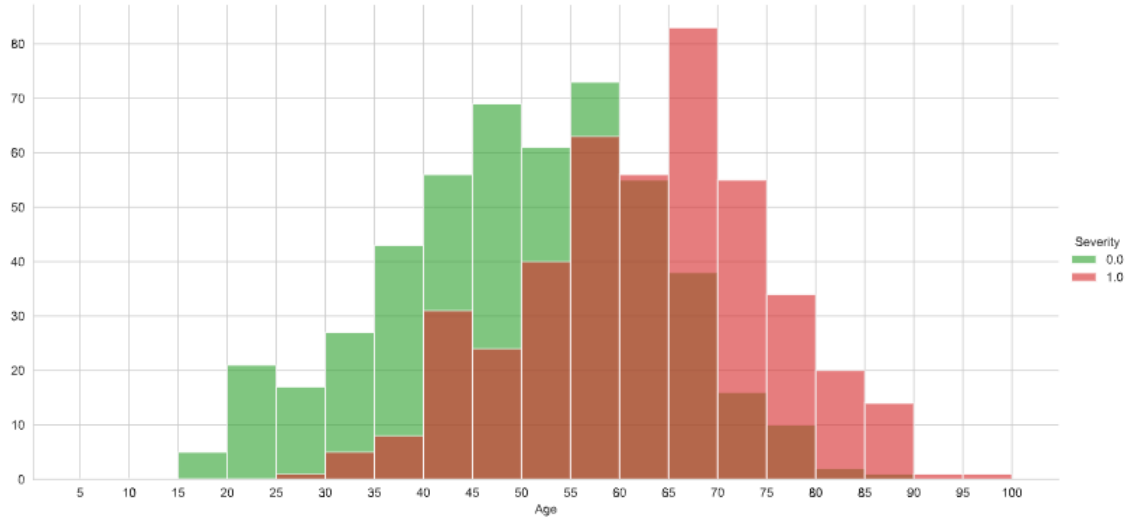

Figura 13: Gráfico de barras das idades

Observando as figuras, podemos observar uma tendência forte no que diz respeito à relação entre as idades mais avançadas e a gravidade do tumor. Verifica-se que as pessoas mais velhas têm uma tendência maior a desenvolver tumores malignos.

Também podemos constatar que a maior parte das nossas amostras estão contidas no intervalo dos 40 e dos 60 anos de idade.

Logo após, passamos à análise da forma do tumor no geral e por classificação de tumores.

mean	2.727957
std	1.249402
min	1.000000
25%	2.000000
50%	3.000000
75%	4.000000
max	4.000000

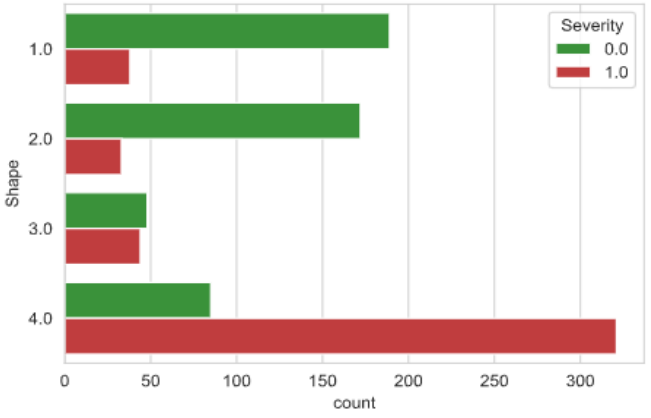

Figura 14: Métricas e gráfico de barras da forma dos tumores

Analisando a figura, é possível confirmar que uma forma **irregular** de um tumor poderá ser um forte indicador de que o tumor é maligno. O contrário acontece quando a forma é **redonda** ou **oval**. Um número bastante superior de amostras para tumores com formas irregulares, é algo que também se deduz.

Prontamente, analisámos a margem do tumor no geral e por classificação de tumores.

mean	2.764516
std	1.567783
min	1.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

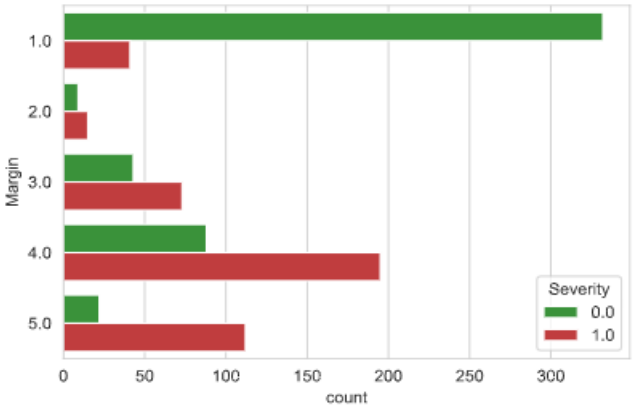

Figura 15: Métricas e gráfico de barras da margem dos tumores

Tendo em conta a figura, é possível constatar que uma margem **circunscrita** é um bom indicador, já que tendencialmente é representativa de tumores benignos. Já uma margem **mal definida** ou **espiculada** pode, ainda que com menos probabilidade, indicar um tumor maligno.

Aqui verifica-se um maior número de amostras para tumores com uma margem circunscrita, seguido de tumores com uma margem mal definida. Infelizmente, são reduzidas as amostras para tumores com margem microlobulada o que nos vai impedir de tirar grandes conclusões.

Seguidamente, analisámos a densidade do tumor no geral e por classificação de tumores.

mean	2.915054
std	0.362816
min	1.000000
25%	3.000000
50%	3.000000
75%	3.000000
max	4.000000

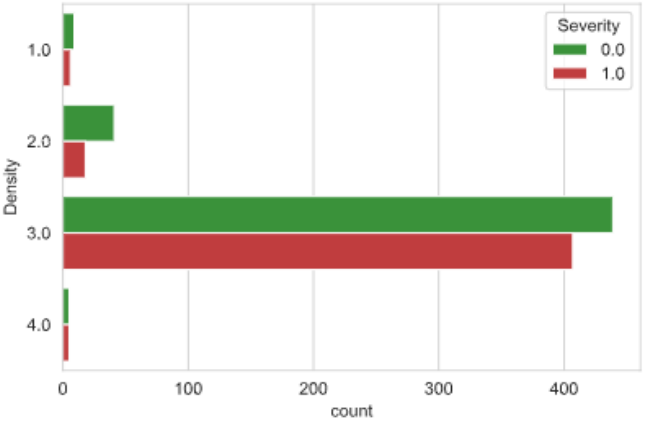

Figura 16: Métricas e gráfico de barras da densidade dos tumores

O facto das amostras estarem todas concentradas numa densidade **baixa** não nos permite avaliar o impacto das outras densidades na classificação final. Ainda para mais, os dados revelam que uma densidade baixa não é determinante para a gravidade do tumor, isto porque a probabilidade de ser ou não maligno é quase a mesma.

Para efeitos de confirmação das conclusões anteriores o grupo desenhou vários gráficos "violino"que relacionam as diferentes *features* com a gravidade do tumor

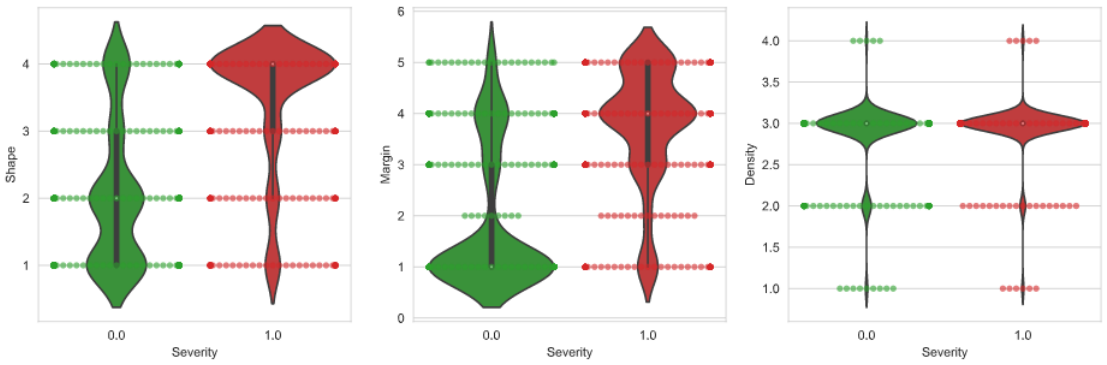

Figura 17: Gráficos "violino"que relacionam as diferentes *features* com a gravidade do tumor

Após uma avaliação da relação das *features* com a classificação de cada objeto, podemos observar as relações entre *features*. Começamos então por analisar a relação da idade com as restantes *features*.

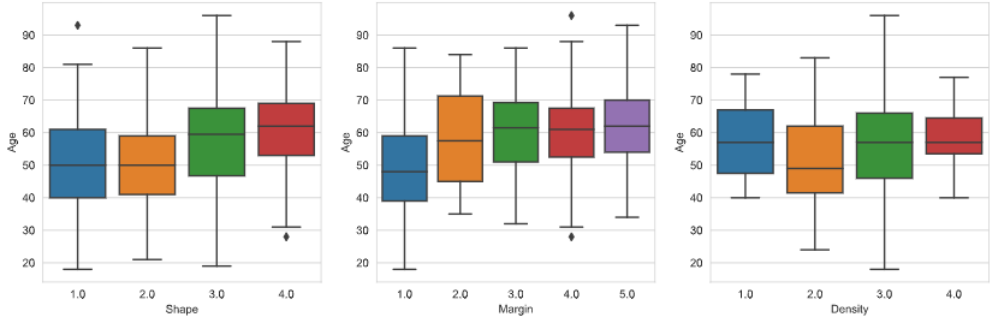

Figura 18: Diagramas de caixa que relacionam as diferentes *features* com a idade

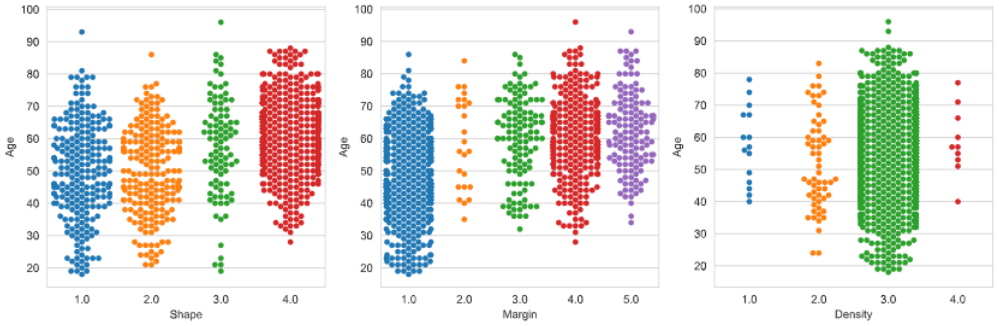

Figura 19: *Swarm-plots* que relacionam as diferentes *features* com a idade

Analisando os gráficos acima, é possível constatar os seguintes factos:

- O valor da caracterização da forma do tumor é diretamente proporcional ao valor da idade. Isto significa que, por exemplo, uma pessoa mais jovem tem tendência a ter tumores com uma forma redonda ou oval, o que, de acordo com as conclusões tiradas da figura 14, nos diz que as pessoas mais jovens terão mais probabilidade de ter tumores benignos.
- O valor da caracterização da margem do tumor é diretamente proporcional ao valor da idade. Isto significa que, por exemplo, uma pessoa mais velha tem tendência a ter tumores com uma margem mal definida ou espiculada, o que, de acordo com as conclusões tiradas da figura 15, nos diz que as pessoas mais velhas terão mais probabilidade de ter tumores malignos.
- Relativamente à densidade, não é possível tirar conclusões relacionadas com a idade, já que praticamente todas as amostras fazem parte de apenas um grupo (densidade = 3.0).

Posteriormente, analisámos a relação entre a margem dos tumores e a forma dos tumores.

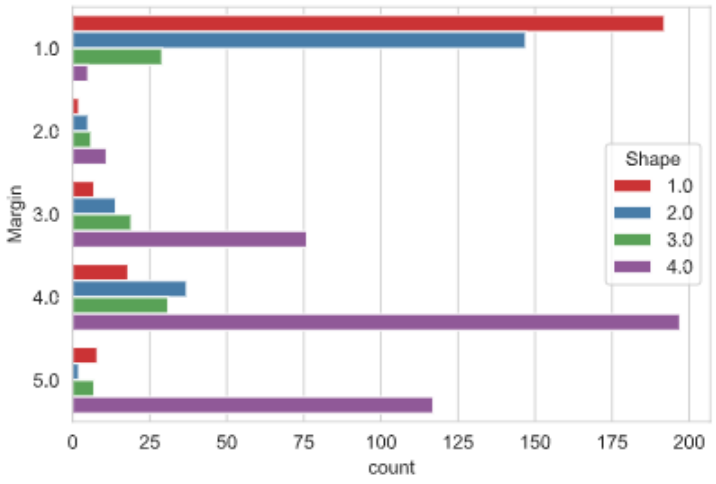

Figura 20: Relação entre a margem e a forma dos tumores

Neste gráfico podemos observar uma relação forte entre a margem e a forma dos tumores. Uma margem circunscrita normalmente indica uma forma redonda ou oval o que, consequentemente, indica uma probabilidade forte do tumor ser benigno. Uma margem mal definida ou espiculada normalmente indica uma forma irregular o que aumenta consideravelmente a probabilidade do tumor ser maligno.

Ainda que as amostras sejam reduzidas para uma margem obscurecida, podemos começar a observar um tendência para uma associação com formas irregular. Sendo assim, podemos estar na presença de um mau indicador.

Prontamente, analisamos a relação entre a densidade e a forma dos tumores.

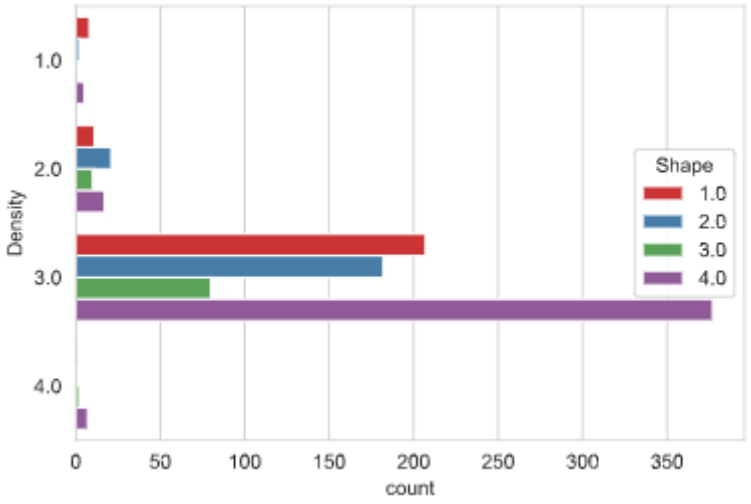

Figura 21: Relação entre a densidade e a forma dos tumores

Aqui não é possível retirar grandes conclusões já que, como já foi referido, a maior parte das amostras tem uma densidade baixa. Podemos observar apenas uma tendência para uma forma irregular quando a densidade é baixa.

Finalmente, investigamos a relação entre a densidade e a margem dos tumores.

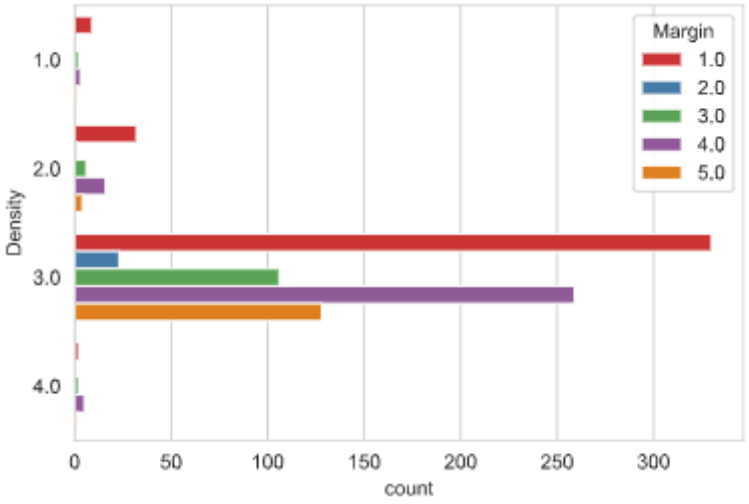

Figura 22: Relação entre a densidade e a margem dos tumores

Aqui podemos verificar que para uma densidade baixa, a margem poderá ser circunscrita ou mal definida o que, na realidade, são indicadores de classificações completamente opostas. Concluímos então que a relação destas duas *features* é extremamente fraca.

5 Processamento dos dados

Após tratamento dos dados em falta e análise de vários gráficos, surge a fase de processamento dos dados, com objetivo de os preparar de forma a potenciar o desempenho do modelo aplicado.

5.1 Dados numéricos

Para dados numéricos, a standardização é uma fase recomendada de pré-processamento para melhorar o desempenho das redes neuronais.

Diferenças nas escalas entre as variáveis de *input* podem aumentar a dificuldade do problema que está a ser modelado. Um exemplo disso são valores de entrada elevados, que pode resultar num modelo que aprende valores com grande peso. Um modelo com grandes valores de peso geralmente é instável, o que significa que pode existir uma quebra de desempenho durante a aprendizagem, resultando num erro maior de generalização. [2]

Um *target* com uma grande variedade de valores pode, por sua vez, resultar em grandes valores de gradiente de erro, fazendo com que os valores de peso mudem drasticamente, tornando o processo de aprendizagem instável. Desta forma, escalar os dados apresenta-se como uma etapa crítica no processamento dos dados no sentido de potenciar o desempenho de qualquer rede neuronal.

Os dados numéricos presentes no *dataset* resumem-se à *feature* `Age`. Devido ao enunciado anteriormente, essa *feature* foi submetida ao processo de standardização, sendo este realizado através da utilização do *StandardScaler* presente no *SkLearn*.

Este processo aplica uma valor nos dados, tornando-os assim standardizados, como mostrado de seguida.

	Age
1	43.0
2	58.0
3	28.0
4	74.0
5	65.0

	Age
1	-0.865910
2	0.164220
3	-1.896040
4	1.263026
5	0.644948

Figura 23: Aplicação do StandardScaler

5.2 Dados categóricos

Para dados categóricos, o *One Hot Encoding* é um processo recomendado para melhorar o desempenho das redes neuronais.

Os dados categóricos são variáveis que contêm valores etiquetados em vez de valores numéricos. O número de valores possíveis geralmente é limitado a um conjunto fixo e cada valor representa uma categoria diferente.

Imensos algoritmos de *Machine Learning* não possuem a capacidade de processar dados de forma categórica. Estes exigem que todas as variáveis de *input* e *output* sejam numéricas. Tal acontece porque os algoritmos são construídos no sentido de poder fornecer um resultado eficiente, e isso só acontece com os dados representados de forma numérica. Isso significa que os dados categóricos devem ser convertidos para o formato numérico. [3]

Uma conversão *one-hot* pode ser aplicada à representação inteira. É nesta fase que a variável codificada inteira é removida e uma nova variável binária é adicionada para cada valor inteiro exclusivo. Desta forma, para cada categoria, vão ser necessárias variáveis binárias quantas as opções existentes, sendo que cada uma das novas colunas vão possuir valor positivo caso esse seja o valor da categoria, e valor negativo caso contrário.

Os dados categóricos do *dataset* que foram submetidos ao processo de *One Hot Encoding* foram as seguintes *features*: `Margin`, `Shape` e `Density`.

De seguida é apresentada a sua aplicação na *feature* `Margin`.

	Margin
1	1.0
2	5.0
3	1.0
4	5.0
5	1.0

	Margin_1	Margin_2	Margin_3	Margin_4	Margin_5
1	1	0	0	0	0
2	0	0	0	0	1
3	1	0	0	0	0
4	0	0	0	0	1
5	1	0	0	0	0

Figura 24: Aplicação do One Hot Encoding

O outro dado categórico presente nos dados é a feature `Severity`, variável reconhecida como *target*. Esta já se encontra no formato binário, e devido a tal não sofreu alterações.

6 Redes Neuronais

O modelo utilizado para resolver o problema apresentado foi uma Rede Neuronal *Feed-forward*, mais especificamente uma *Multilayer Perceptron*. A construção desta foi realizada com o software *Keras*, que funciona com base no *Tensorflow*.

A rede apresenta 14 nodos de *input* que são, especificamente: a `Age`, as 5 hipóteses da `Margin`, as 4 hipóteses da `Shape` e as 4 hipóteses da `Density`. Esta apresenta um único nodo de output, neste caso o *target* `Severity`. A camada de entrada possui a função de ativação `sigmoid`, assim como a camada de saída. Deste modo, garantimos que os valores são diferenciados de forma binária logo na primeira camada e que os valores da camada de saída se encontram compreendidos entre 0 e 1. Todos os restantes parâmetros da arquitetura da rede são selecionados de forma otimizada por um algoritmo genético, apresentado na secção seguinte.

Na seguinte representação da rede são apresentadas duas camadas intermédias, no entanto essa é apenas uma das possibilidades das quais o algoritmo genético pode optar.

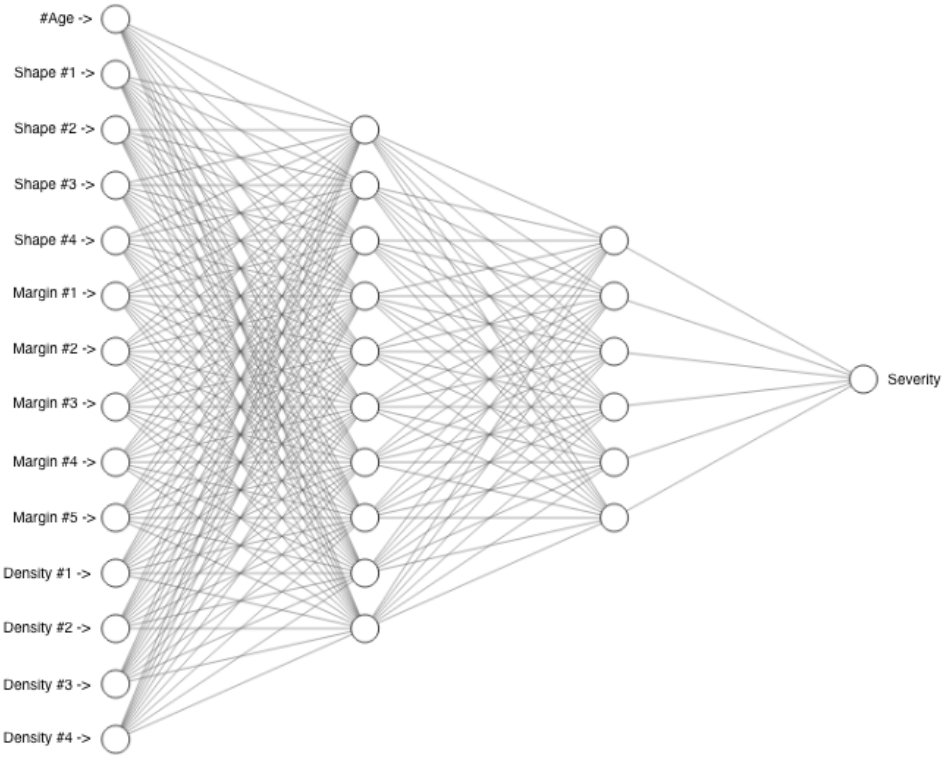

Figura 25: Arquitetura típica da rede

De forma a evitar *overfitting* com o conjuntos de dados no qual está a ocorrer o treino foram utilizadas técnicas de *Cross Validation*. Sem esta aplicação teríamos informações ajustadas sobre o desempenho do modelo nos dados da amostra, o que seria tendencioso e negativo para futuras previsões do modelo.

7 Algoritmos Genéticos

Após criada uma primeira arquitetura da rede neuronal, está inerente a procura pelos parâmetros que proporcionam o melhor desempenho. Neste sentido, surgem os algoritmos genéticos que assumem a função de encontrar a conjugação de parâmetros que melhor se adequam ao problema.

Um algoritmo genético é uma heurística de pesquisa inspirada na teoria da evolução natural de Charles Darwin. Este algoritmo reflete o processo de seleção natural, onde os cromossomas mais aptos são selecionados para reprodução, a fim de produzir descendentes da próxima geração. Um dos principais processos neste algoritmo é a seleção natural. Este começa com a seleção dos cromossomas mais aptos de uma população, que produzem "filhos"que herdam as características dos "pais"e são adicionados à próxima geração. Se os "pais"tiverem melhores resultados, os "filhos"serão melhores que os "pais"e terão ainda melhores resultados. Este processo é repetido até que, no final, seja encontrada uma geração com os cromossomas mais aptos. [4]

Esta noção pode ser aplicada a um problema de pesquisa, sendo que é esta a vertente que faz sentido ser aplicada à problemática apresentada no início desta secção. Objetivamente, consideramos um conjunto de soluções para o problema e selecionamos a melhor.

O algoritmo genético procede por diversas fases, sendo cada uma destas abordadas nas seguintes subsecções.

7.1 Constituição dos cromossomas

O processo do algoritmo genético começa com um conjunto de cromossomas denominado população, no qual cada cromossoma é uma solução para o problema em questão. Cada um dos cromossomas é caracterizado por um conjunto de parâmetros variáveis denominados de genes. Os genes são compactados de modo a formar um cromossoma (solução).

Para o problema em causa, os genes utilizados foram os parâmetros referentes à rede neuronal. Esse parâmetros são os que se adequam ao problema em causa, neste caso, uma classificação com *target* binário. Os parâmetros possíveis são os seguintes:

- **Número de camadas intermédias**, com um valor possível entre `1` e `16`;
- **Número de nodos por camada intermédia**, com um valor de entre os seguintes: `1, 2, 4, 8, 16, 32, 64, 128, 256`;
- **Função de ativação**, de entre as seguintes: `relu, selu, sigmoid, tanh, linear, softmax`;
- **Taxa de aprendizagem**, com um valor de entre os seguintes: `0.001, 0.01, 0.1, 1`;
- **Função de otimização**, de entre as seguintes: `SGD, RMSprop, Adam`;
- **Função de perda**, de entre as seguintes: `binary_crossentropy, hinge, squared_hinge`;

7.2 Avaliação dos cromossomas

Após escolha, numa primeira fase aleatória, dos parâmetros incluídos nos cromossomas, uma rede neuronal é treinada com cada um desses conjuntos de parâmetros.De forma a produzir uma maior diversidade de resultados e enriquecer o processo de validação da rede, cada cromossoma é validade com *K-Fold Cross Validation*, aplicado sobre o conjunto de dados de treino, previamente separado através da função `test_train_split`.

Dado o contexto do problema, procuramos elaborar uma métrica mais expressiva que simplesmente a *accuracy*. Geralmente, em previsões enquadradas em contexto médico, é importante procurar minimizar o número de falsos negativos (casos que o modelo classificador desenvolvido classificaria como sendo cancros benignos, mas, na verdade, seriam tumores malignos). Assim, de cada *fold* recolhemos as seguintes métricas, comparando os valores previstos com os valores verdadeiros:

- *accuracy*
- *recall* por classe
- *precision* por classe

No final dos dez *folds* (k=10), é feita uma média de cada um deste valores. O *score* de um determinado cromossoma é determinado pela seguinte expressão:

$$score = 0.5 * accuracy + 0.175 * recall_1 + 0.15 * precision_0 + 0.1 * recall_0 + 0.075 * precision_1 \quad (1)$$

Cada resultado é comparado, sendo que a probabilidade de um cromossoma ser selecionado para reprodução é baseado seu *score*.

7.3 Seleção

A ideia da fase de seleção é selecionar os cromossomas com melhores resultados e permitir que eles passem seus genes para a próxima geração. Deste modo, os **cinco** cromossomas com melhor *score* são selecionados para prosseguirem para a próxima geração.

7.4 Crossover

Esta é a fase mais importante de um algoritmo genético. Para cada conjunto de cromossomas "pai", um ponto de cruzamento é escolhido aleatoriamente de entre os genes. Os "filhos"são criados trocando os genes dos pais entre si até que o ponto de cruzamento seja alcançado e, depois, adicionados à população.

7.5 Mutação

Nos cromossomas "filho"formados, alguns dos genes são submetidos a uma mutação aleatória. Isto representa que alguns dos parâmetros dos conjuntos indicados anteriormente são alterados. Tal acontece de forma a manter a diversidade dentro da população e impedir uma convergência prematura.

7.6 Desempenho do Algoritmo Genético

O objetivo do algoritmo genético é realizar uma otimização automática da arquitetura da rede, isto é, dos parâmetros aplicados na rede. De forma a poder visualizar a evolução do *score* com o decorrer das gerações, é apresentado o seguinte gráfico onde se verifica a melhoria dessa métrica com a execução do algoritmo.

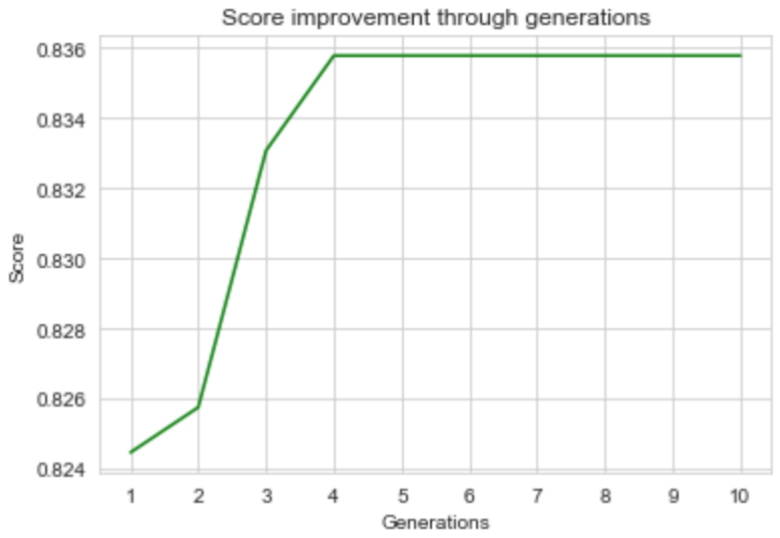

Figura 26: Gráfico demonstrativo da evolução do Algoritmo Genético

7.7 Resultado obtido

Do processo de aplicação do algoritmo genético, os parâmetros selecionados foram os seguintes:

- Número camadas intermédias: 14
- Nodos por camada: 256
- Função de ativação: linear
- Otimizador: SGD
- Item Aprendizagem: 0.01
- Função de *loss*: hinge

Dada a natureza aleatória associado ao processo de K-fold Cross Validation, assim como a divisão e *shuffle* do *dataset* realizados cada vez que se repete este processo, os parâmetros escolhidos variam.

Para concluir o *pipeline*, resta avaliar o desempenho de uma rede neuronal parametrizada de acordo com o resultado do algoritmo genético. Assim, fazemos *fit* dos dados de teste num novo modelo, treinado ao longo de 50 épocas. Em cada época, 1/3 do conjunto de dados foi utilizado para validação (*validation_split=0.33*). Os resultados e a evolução deste treino são sumarizados nas imagens que se seguem:

Figura 27: Evolução da *loss* ao longo das épocas

Figura 28: Evolução da *accuracy* ao longo das épocas

Por fim, e atendendo ao facto deste modelo ter sido gerado tendo por base mais métricas além da *accuracy*, podemos avaliar o resultado obtido através de uma matriz de confusão e das métricas que foram utilizadas no processo:

```
              precision    recall  f1-score   support

      Benign       0.84      0.85      0.85       159
   Malignant       0.80      0.79      0.80       121

    accuracy                           0.82       280
   macro avg       0.82      0.82      0.82       280
weighted avg       0.82      0.82      0.82       280

True Positives: 96
True Negatives: 135
False Positives: 24
False Negatives: 25
```

Figura 29: Resultados obtivos

Analisando os resultados, podemos verificar que conseguimos classificar corretamente 82% dos casos. No entanto, o modelo ainda apresentou 25 falsos negativos, um valor que pode ser otimizado.

8 Segunda abordagem

Após aplicar os métodos apresentados, surgiu outra abordagem com intenção de diferenciar mais os dados e tornar mais real a informação obtida do *dataset*:

- Remoção de todas as linhas que contêm valores nulos, de forma a todos os dados contidos no *dataset* serem reais;
- Remoção da *feature* `Densidade`, por ser predominantemente de valor 3 e possuir uma taxa de valores nulos na ordem dos 7%, consideravelmente alta;
- Remoção de outliers, através da função *Z-Score*;
- Discretização da idade, com os seguintes valores: `<40, 40 - 55, 55 - 60, >60`;

8.1 Outliers

De forma a não introduzir dados que se encontrem fora do estado habitual no modelo, foi realizada uma procura por *outliers* nos dados, da função matemática do *Z-Score*.

Para cada coluna, primeiro é calculada o `Z-Score` de cada valor na coluna, em relação à média e desvio padrão da coluna, e considerado o valor absoluto da pontuação. Após estes cálculos, é considerado um *threshold* e apenas são mantidos todos os valores abaixo desse limite. No exemplo aqui apresentado foi considerado um *threshold* de valor 2, no qual são considerados 95% dos valores que se encontram mais perto da média. Desta forma, garantimos que não existem exceções a enviesar o normal funcionamento do modelo.

8.2 Discretização da característica Idade

Uma vez que todas as características do *dataset*, à exceção da idade, são dados categóricos, procuramos também transformar a idade num. Assim, procedemos à sua discretização em 4 *bins*, estabelecidos tendo em conta o *swarm-plot* da figura 12. Para criação destes *bins*, procurou-se identificar intervalos de idade que exibissem uma determinado padrão na severidade do tumor. O resultado deste processo é representado de seguida:

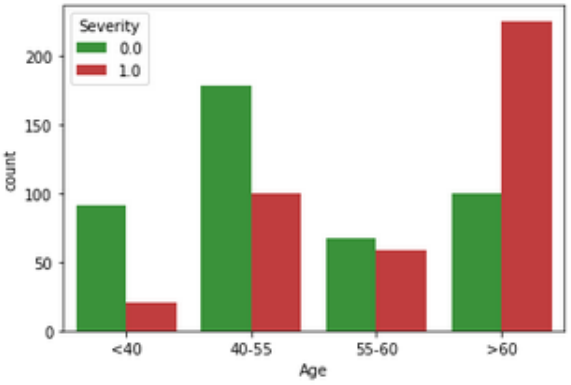

Figura 30: Gráfico de barras com as idades discretizadas

Neste gráfico verifica-se a existência duma população de baixo risco, isto é, as pessoas com menos de 40 anos. De seguida, a zona dos 40 aos 55 na qual os casos negativos já assumem uma preponderância maior. Entre os 55 e os 60 os cancros já aparecem com tanta regularidade como os que não aparecem, e por fim, nos doentes com mais de 60 anos os tumores são maioritariamente malignos. Uma vez findo este processo, aplicamos o processo de *One Hot Encoding* a esta *feature* também.

8.3 Resultados obtidos

Face às experiências que verificamos ao longo do processo até então, esta abordagem parecia promissora na medida em que, agora, todas as caraterísticas seriam categóricas, o que poderia trazer melhorias a nível da escolha da função de ativação e de perda. Além disso, ao descartamos a característica Densidade, estaríamos a simplificar o modelo e a remover uma *feature* com pouca influência na severidade do tumor. No entanto, os valores registados ao longo de diferentes testes não conseguiram ultrapassar os que foram registados pela primeira abordagem.

9 Conclusão

Após a realização deste trabalho conseguimos, de facto, compreender os resultados que obtivemos, tendo em conta as decisões tomadas ao longo do projeto. São exatamente estes resultados e estas decisões que, nesta conclusão, vamos analisar e justificar, respetivamente.

A análise inicial foi algo banal. A único ponto a abordar terá sido a remoção da coluna **BI-RADS** que, como já foi explicado, não representa algo importante para a classificação de tumores.

Na fase de tratamento de valores nulos, o grupo optou por remover as linhas com 2 ou mais *missing values* e manter aquelas que tivessem apenas 1 *missing value*, substituindo o valor desse *missing value* pela moda correspondente à classe associada a este. Esta estratégia foi adotada na tentativa de manter o maior número de dados possíveis, sem os enviesar, através da sua manipulação.

De seguida, ocorreu a fase de EDA (*Exploratory Data Analysis*). Esta fase serviu, principalmente, para nos familiarizarmos ainda mais com os dados em questão. Sendo assim, não há grandes decisões a tomar pelo que não é necessário justificar nem analisar qualquer tipo de decisão.

A fase seguinte, correspondente à fase de *Data Preparation*, é extremamente importante. O *One Hot Encoding* e a Standardização representam processos importantíssimos no que diz respeito à eficiência de algoritmos de Redes Neuronais e, como este modelo é constituído por uma RNA, tomamos também a decisão de os realizar.

No que diz respeito ao processo de otimização da rede neuronal, aplicamos um algoritmo genético cuja mutação entre gerações se baseada numa métrica personalizada, desenvolvida tendo em especial atenção o contexto do problema, com o objetivo de procurar minimizar o número de falsos negativos. Além disso, foram explorados e estudados a maior parte dos parâmetros através dos quais é possível configurar uma rede neuronal, com recurso ao Keras.

Analisando os resultados obtidos, podemos afirmar que estes são proporcionais ao esforço envolvido na realização deste trabalho, estando a par de resultados obtidos como em [5] e [6]. Certamente a aplicação do processo de *One Hot Encoding* terá sido fulcral para atingir os valores apresentados, razão que motivou a segunda abordagem descrita no capítulo 8. No entanto, discretizar a idade e remover a caraterística Densidade não provocou um aumento substancial no desempenho do modelo.

As principais dificuldades sentidas passaram pela aprendizagem de todos os parâmetros que podem ser utilizados para configurar a rede neuronal, sendo que foi realizado um estudo prévio, com alguma profundidade, que nos permitiu reduzir o âmbito do espaço de procura de soluções àquelas que melhor se adequam para o tratamento de dados categóricos, como os presentes no *dataset*.

Como trabalho futuro iremos avaliar a performance do processo de otimização aplicado, comparando-o com outros, como, por exeplo, o *Grid Search*. Seria também interessante avaliar os tempos necessários para otimizar a rede utilizando como backend o TensorFlow GPU, contrapondo-os com a versão utilizada (CPU).

Referências

[1] "Óbitos por algumas causas de morte (%)". https://www.pordata.pt/Portugal/óbitos+por+algumas+causas+de+morte+(percentagem)-758

[2] Jason Brownlee. "How to use Data Scaling Improve Deep Learning Model Stability and Performance", Fev 2019. https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

[3] Jason Brownlee. "Why One-Hot Encode Data in Machine Learning?", Jul 2017. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[4] Vijini Mallawaarachchi. "Introduction to Genetic Algorithms", Jul 2017. https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3

[5] Kaushik, Divyansh, and Karamjit Kaur. Application of Data Mining for High Accuracy Prediction of Breast Tissue Biopsy Results.

[6] Lairenjam, Benaki, and Siri Krishan. "A Note on Analysis of Mammography Data." Int. J. Open Problems Compt Math, vol. 3, no. 5, 2010, www.emis.de/journals/IJOPCM/Vol/10/IJOPCM(vol.3.5.4.D.10).pdf.