Aplicação de MLP

EQUIPE :Leonardo Sena, Pedro Gustavo Santana e Vitor Carvalho Pontes.
Disciplina: Inteligência computacional.
Docente: Anusio Menezes.

Este documento tem como objetivo exibir e explicar as duas base de dados retiradas do UCI Machine Learning Repository e mostrar a aplicação delas em uma rede neural com *tensorflow* e *keras*.

A primeira base informa sobre as características do câncer de mama de pacientes e classifica por pacientes com câncer reincidente ou não. O problema proposto para a rede ao ser treinada por esses dados é se ela seria capaz de prever se um paciente poderia ter reincidência de câncer. Já a segunda base de dados traz todas as possíveis jogadas que podem acontecer pelo jogador no jogo TIC TAC TOE Endgame, popularmente conhecido como jogo da velha. O problema proposto para a rede ao ser treinada por esses dados é se ela seria capaz de informar se o jogador X ganharia de acordo com um dada configuração de campos.

Experimentos

Câncer de mama

Os dados fornecidos pelas universidades University Medical Centre (Londres) e Institute of Oncology Ljubljana (Eslovênia), Yugoslavia são separados 9 atributos de caracterização e 1 de classificação, com um total 286 instâncias, sendo os atributos e a classificação:

- Idade;
- Menopausa;
- Tamanho de tumor;
- Invasão linfonodal;
- Nódulo capsulado;
- Grau histológico do tumor:
    - Predomínio de células cancerígenas;
    - Consiste em características usuais de células cancerígenas;
    - Células altamente afetadas.
- Mama (lado);
- Mama (quadrante);
- Histórico de radioterapia.
    - Recorrente ou não recorrente. Pessoas que voltaram a ter o câncer de mama.

Abaixo será apresentado os dados da base de dados em formato de gráfico. O gráficos mostram a quantidade de pacientes por um atributo de caracterização (idade, tamanho do tumor, etc.).

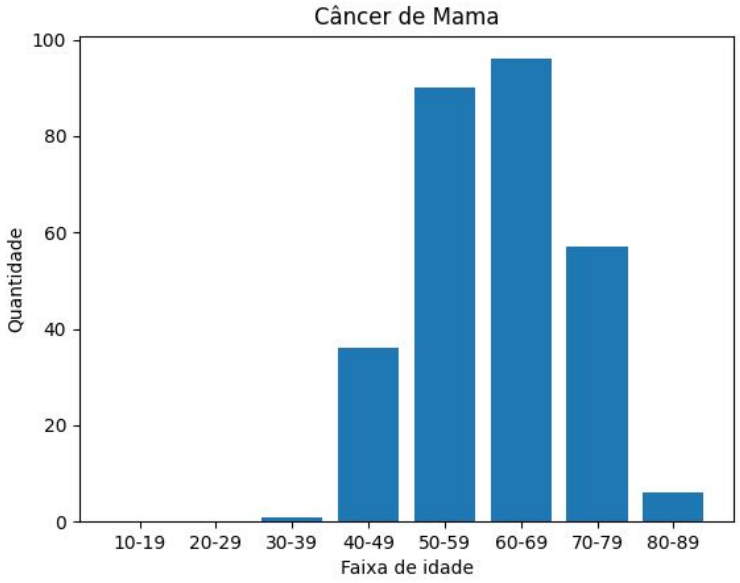

Figura 1: Gráfico que apresenta a quantidade de pacientes por idade.
Fonte: Própria.

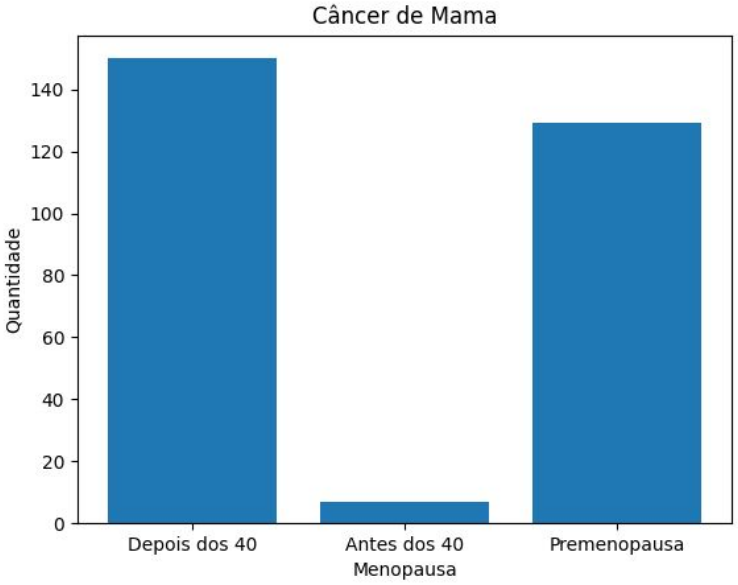

Figura 2: Gráfico que apresenta a quantidade de pacientes por período da menopausa.
Fonte: Própria.

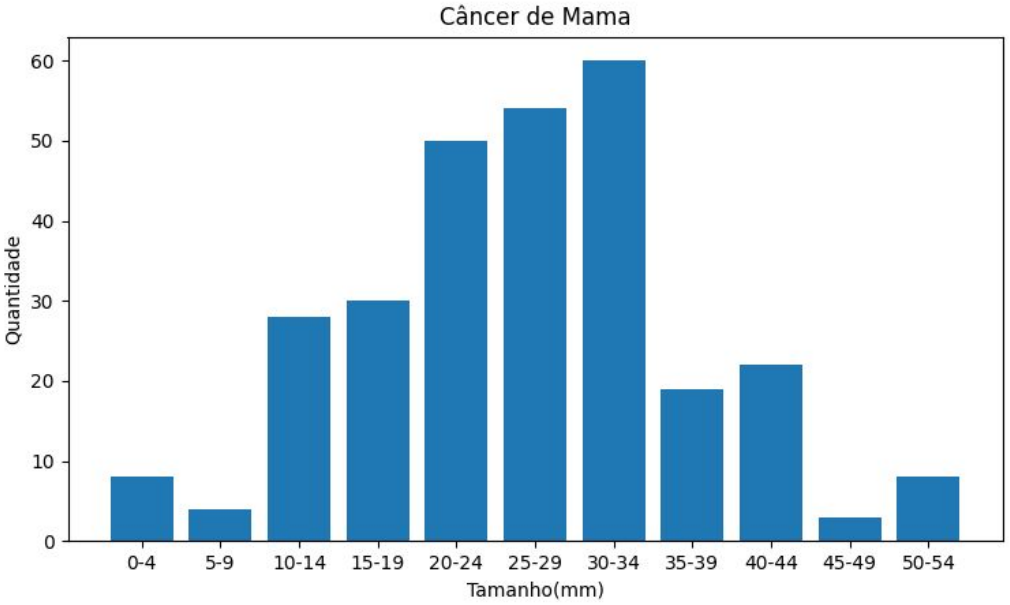

Figura 3: Gráfico que apresenta a quantidade de pacientes por tamanho do câncer.
Fonte: Própria.

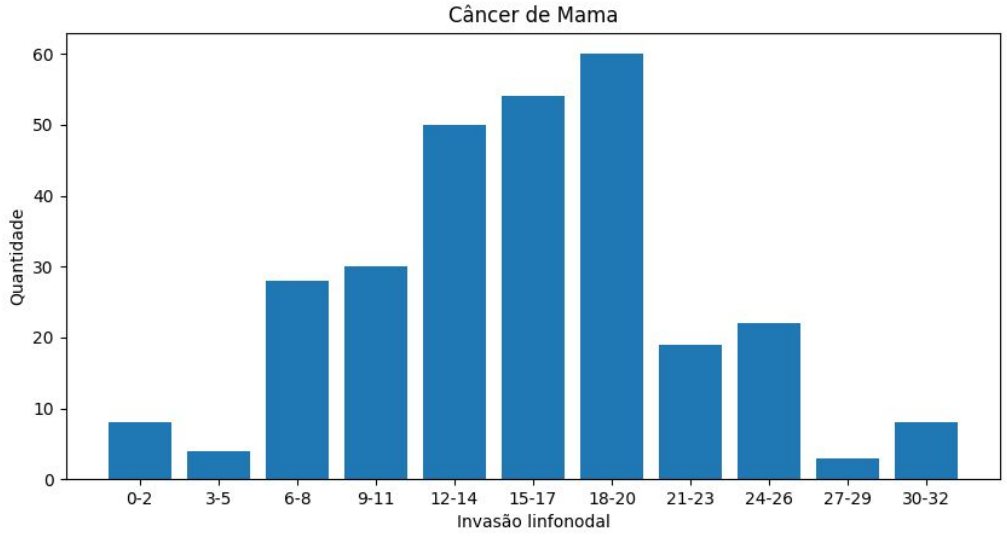

Figura 4: Gráfico que apresenta a quantidade de pacientes por quantidade de linfonodos invadidos.
Fonte: Própria.

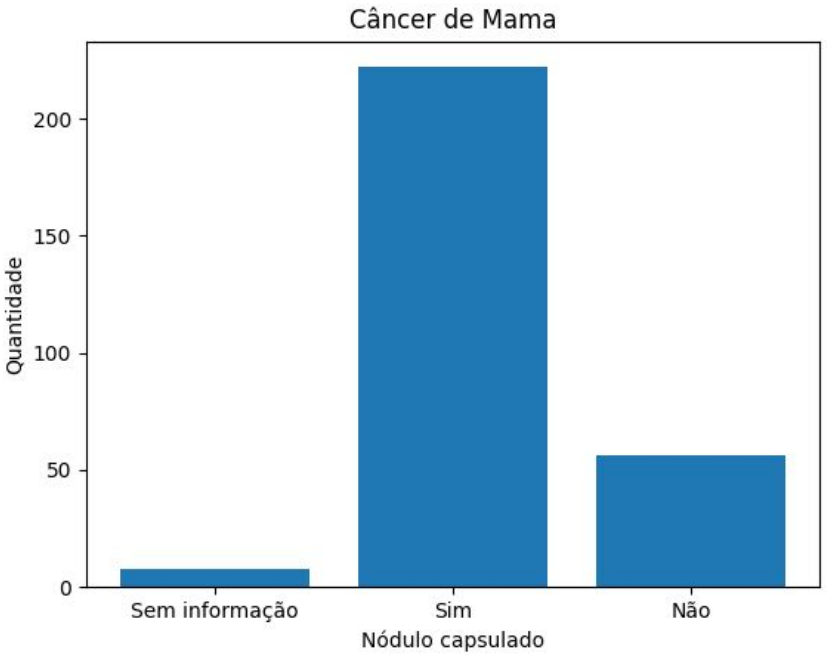

Figura 5: Gráfico que informa quantidade de pacientes por tipo de encapsulamento do tumor.
Fonte: Própria.

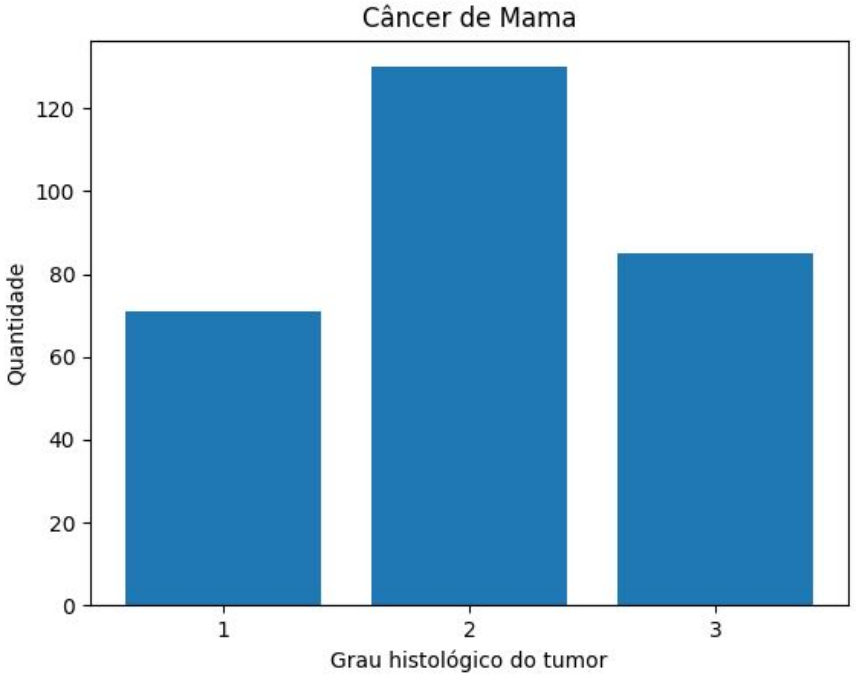

Figura 6: Gráfico que informa quantidade de pacientes por forma histológica do tumor.
Fonte: Própria.

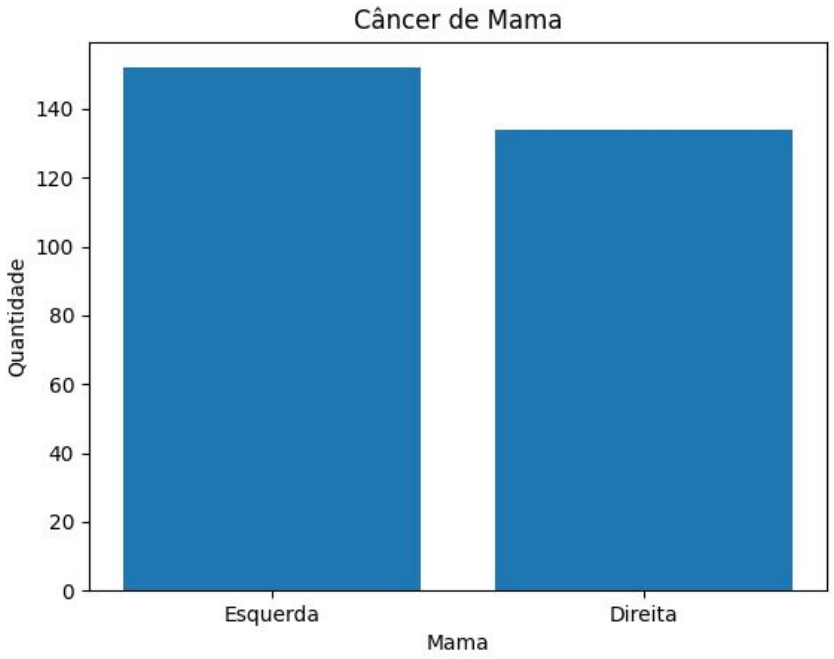

Figura 7: Gráfico que informa quantidade de pacientes de acordo qual mama está localizado o tumor.
Fonte: Própria

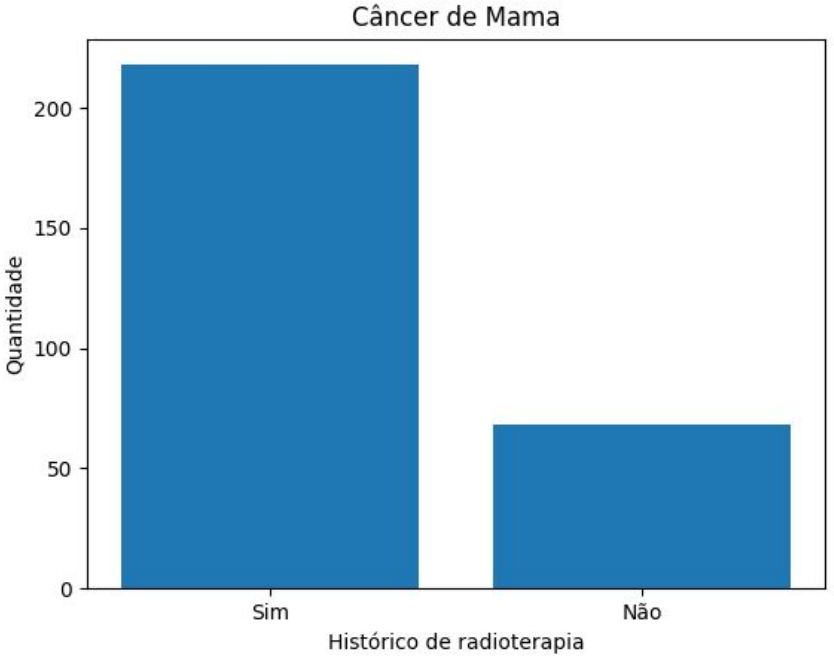

Figura 8: Gráfico que informa quantidade de pacientes por histórico de radioterapia.
Fonte: Própria.

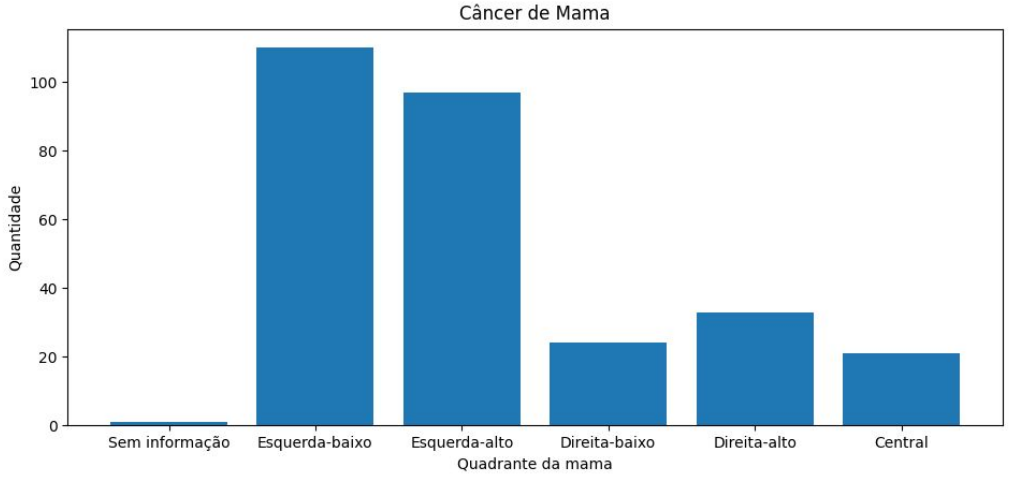

Figura 9: Gráfico que informa quantidade de pacientes por quadrante da mama.
Fonte: Própria.

Após estudo dos gráfico é possível concluir que a grande maioria dos pacientes optaram pela radioterapia, sendo maior parte desses números pessoas que estão na pré menopausa e pessoas que possuem mais de 40 anos. Os quadrantes com maior presença de tumor foram o esquerdo-baixo e o esquerdo-alto.

Abaixo será apresentado a disposição dos dados originais da base de dados. Percebe-se que 5 das características dos dados (idade, menopausa, invasão nodal, grau de histológico e mama) são assimétricos. Assim, como a presença de *outliers*, pontos muitos distantes no quadrantes.

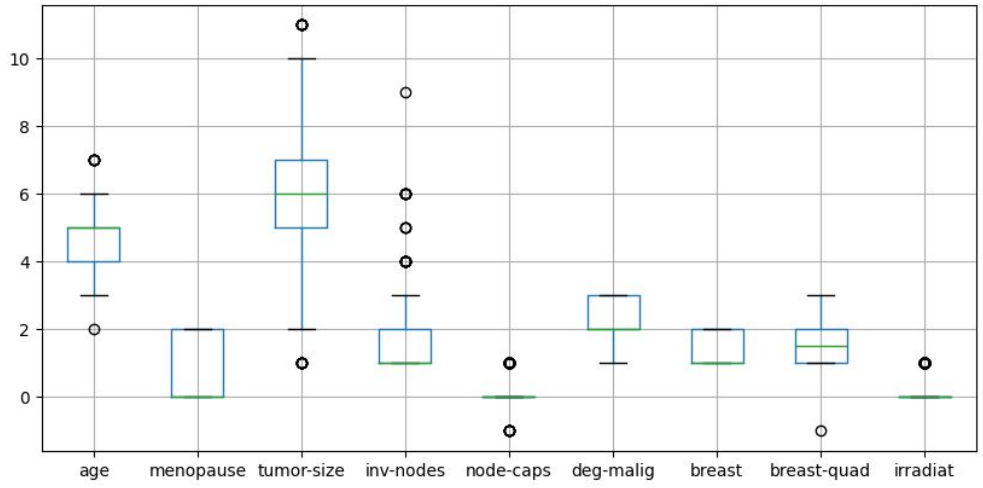

Figura 10: Gráfico *boxplot* que apresenta a mediana, máximos, mínimos, os quartis e *outliers* da base de dados originais. Fonte: Própria..

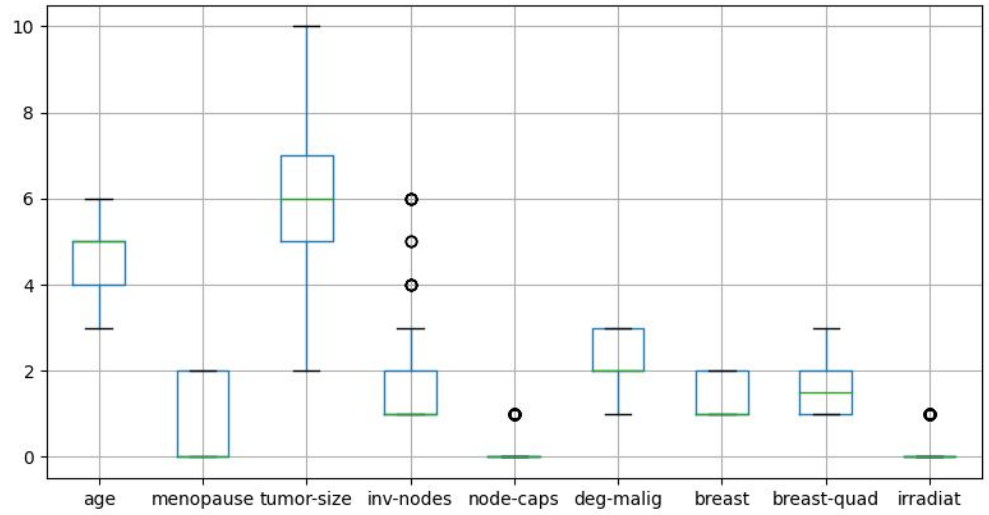

Figura 11: Gráfico *boxplot* que apresenta a mediana, máximos, mínimos, os quartis e *outliers* da base de dados adaptada para o experimento. Fonte: Própria.

Para a utilização dessa base foi necessário fazer alterações, não sendo necessário a alteração do grau histológico do tumor. Ajustes para utilização:

- Idade:
    - 30 - 39 = 3;
    - 40 - 49 = 4;
    - 50 - 59 = 5;
    - 60 - 69 = 6;
    - 70 - 79 = 7.
- Menopausa:
    - Premeno = 0;
    - Lt40 (Antes) = 1;
    - Ge40 (Depois) = 2.
- Tamanho de tumo:
    - 0-4 = 1;
    - 5-9 = 2;
    - 10-14 = 3;
    - 15-19 = 4;
    - 20-24 = 5;
    - 25-29 = 6;
    - 30-34 = 7;
    - 35-39 = 8;
    - 40-44 = 9 ;
    - 45-49 = 10;
    - 50-54 = 11;
    - 55-59 = 12.
- Invasão linfonodal:
    - 0-2 = 1;
    - 3-5 = 2.
    - 6-8 = 3;
    - 9-11 = 4;
    - 12-14 = 5;
    - 15-17 = 6;
    - 18-20 = 7;
    - 21-23 = 8;
    - 24-26 = 9;
    - 27-29 = 10;
    - 30-32 = 11;
    - 33-35 = 12;
    - 36-39 = 13.
- Nó capsulado:
    - Não = 0;
    - Sim = 1.
- Mama (lado):
    - Esquerdo = 1;
    - Direito = 2.
- Mama (quadrante):
    - Esquerdo baixo = 1;
    - Esquerdo alto = 1.5;
    - DIreito baixo = 2;
    - Direito alto = 2.5.

- Histórico de radioterapia:
  - Não = 0;
  - Sim = 1

Outros ajustes necessários foram a retirada de linha que não possuíam uma característica e retirada de outliers (o máximo possível, pois alguns a retirada de alguns outliers interferiam de forma significativa na base de dados). Ao final, o total foi de 256 instâncias. A base de dados foi dividida em 3 partes, a primeira parte (entrada de treinamento) são os dados de treinamento (os primeiros 85), a segunda parte são os dados de validação do treinamento (os 85 subsequentes) e terceira parte são os dados de predição (os últimos 85), distribuídos da forma mais uniforme possível.

Foi utilizado perceptron de multicamada (MLP) com 4 camadas escondidas (12, 8, 4 e 1 node), com a função de ativação 'softmax' nas 3 primeiras e na última foi utilizado 'relu', com 'adamax' como otimizador da rede, para os treinamentos épocas, quantidade de *batch size* e 'mean_squared_logarithmic_error' como parâmetro de perda. Foi atribuído ao primeiro treinamento 200 épocas e um *batch size de 16*, obtendo como resultado uma taxa de acerto médio na rede de treinamento de 77,81% com um erro médio de 0,07, já para o segundo treinamento foi atribuído 125 época e um *batch size* de 64 com o *validation data* (dados de validação separados da base de dados), obtendo como resultado uma taxa de acerto médio de 88,27% e um erro de 0,05. A evolução do acerto e do erro podem ser acompanhados pelos gráficos abaixo do treinamento 1 e 2, respectivamente.

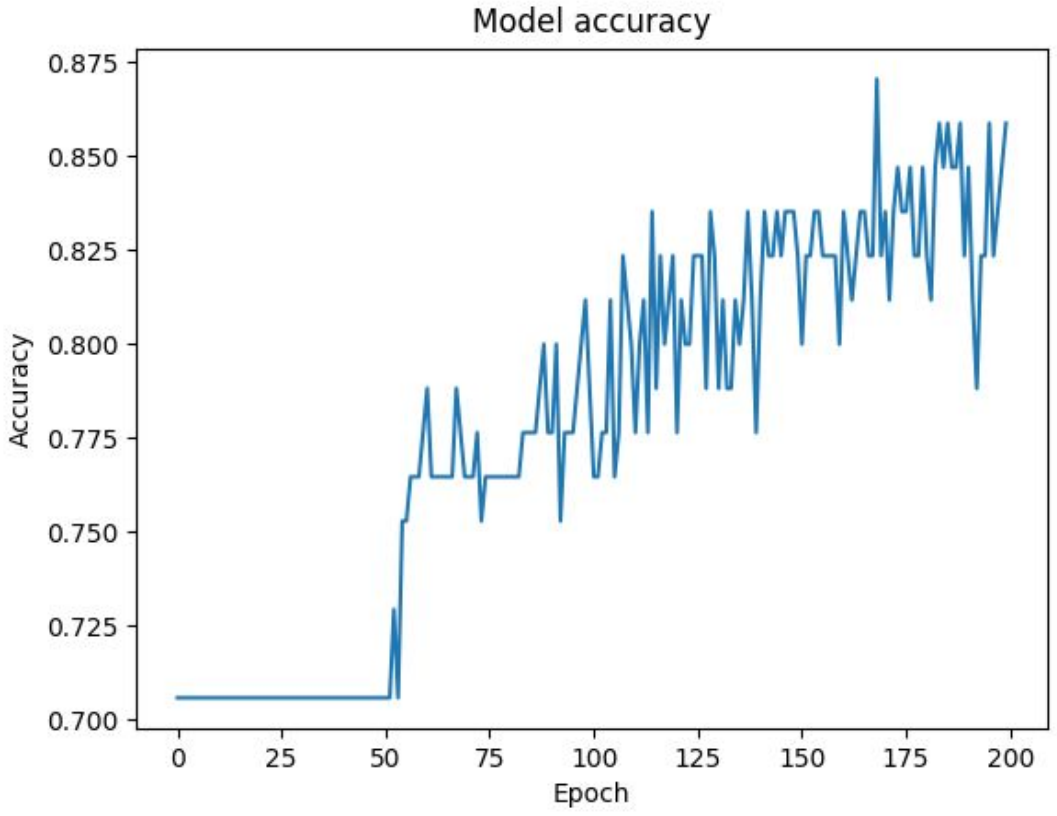

Figura 12: Gráfico demonstrativo do modelo accuracy contendo a evolução dos acertos.Fonte: Própria.

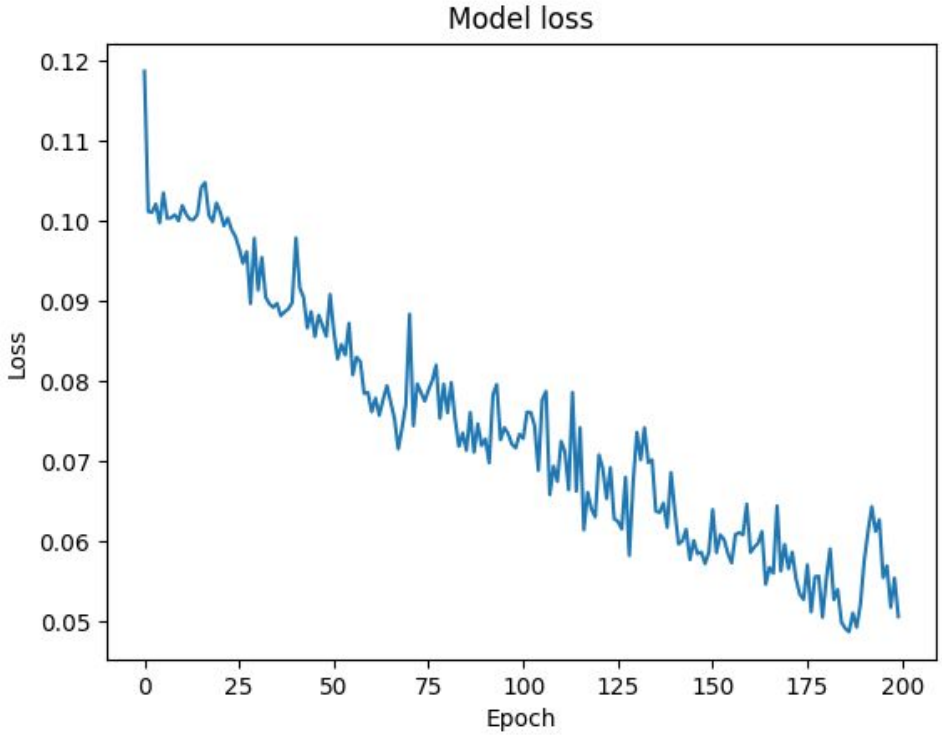

Figura 13: Gráfico demonstrativo do model loss contendo a evolução dos erros.Fonte: Própria.

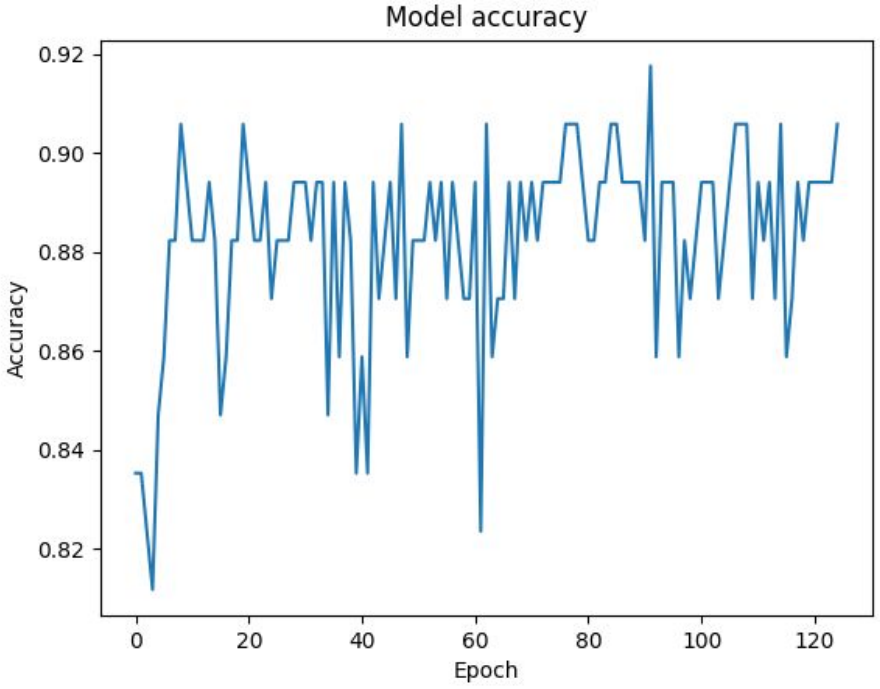

Figura 14: Gráfico demonstrativo do modelo accuracy contendo a evolução dos acertos.Fonte: Própria.

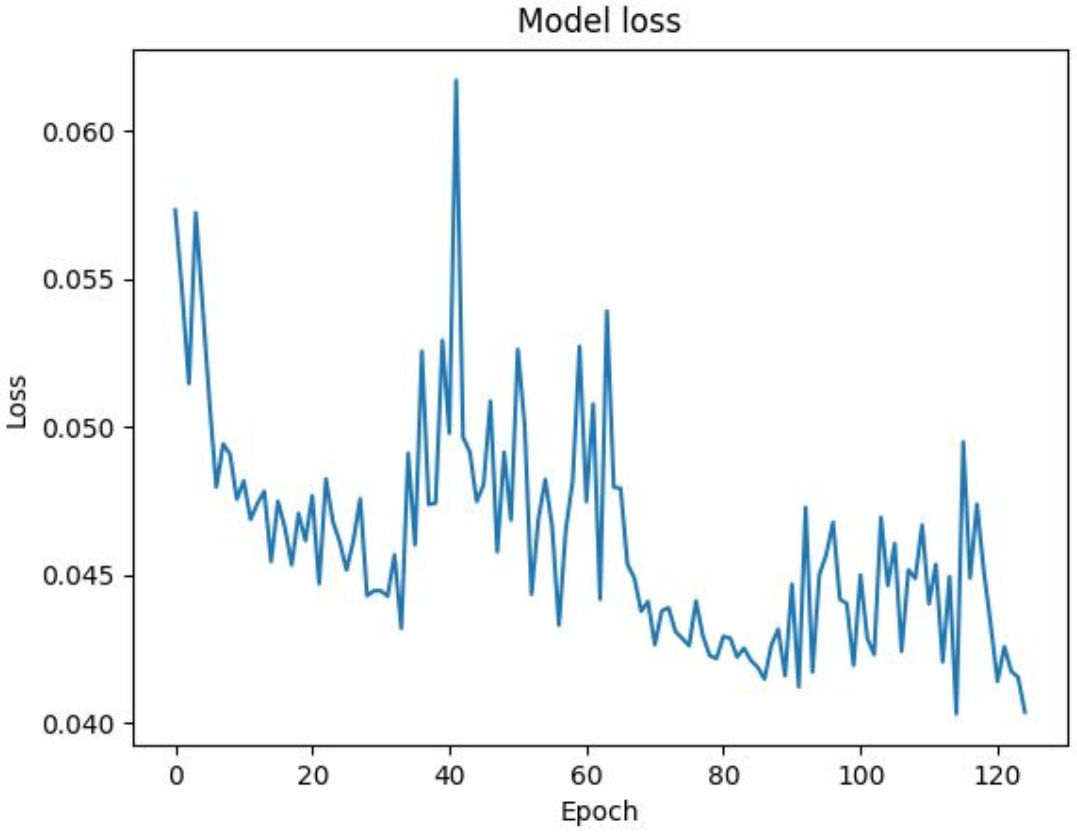

Figura 15: Gráfico demonstrativo do model loss contendo a evolução dos erros.Fonte: Própria.

Após os treinamentos, a rede passou por uma validação com os dados de validação obtendo um um acerto de 74,42% e um erro de 0.11.

Os hiperparâmetros citados foram escolhidos depois da rede apresentar os melhores resultados no treinamento. Na avaliação da rede, com dados de predição a rede teve um acerto de 39%. Apesar dos bons resultados apresentados no treinamento, quando submetida a outro conjunto de dados, que nunca foram vistos antes, a rede não apresentou resultados satisfatórios. Sendo assim, não podendo ser aplicada para tentar indicar se o paciente irá ter reincidência do câncer de mama.

Apesar dos bons resultados apresentados no treinamento, quando submetida a outro conjunto de dados, a rede não apresentou resultados satisfatórios. Sendo assim, não podendo ser aplicada para tentar indicar se o paciente irá ter reincidência do câncer de mama.

Tic Toc TOE

Na segunda base existem 958 instâncias que possuem 9 atributos de caracterização e 1 de classificação. Sendo eles:

- Esquerdo superior;
- Meio superior;
- Direito superior;
- Esquerdo meio;
- Meio meio;
- Direito meio;
- Esquerdo inferior;
- Meio inferior;
- Direito inferior;
- Positivo ou Negativo.

A base de dados TicTacToe informa todas as possibilidades de jogo, assumindo que o jogador “X” tenha iniciado a partida. Nela caso exista a capacidade do jogador X ganhar a partida o indicativo de classificação será positivo, caso contrário, negativo.

Para a utilização dessa base foi necessário fazer alterações, conforme tabela abaixo:

Valor real	Novo valor substituído
X	1
O	0
B	2
Positive (classificação)	1
Negative (classificação)	0

Tabela 1: Alterações realizadas para utilização da database.
Fonte: Própria.

A partir de todas as vitórias obtidas no jogo, e analisando apenas elas, é perceptível verificar as posições em que o há maiores êxitos. Na figura a seguir é possível perceber que quando o jogador marca as posições TL(Top Left), TR(Top Right), MM(Meio Meio), BL(Base Left) e BR (Base Right) a probabilidade de alcançar a vitória é maior que nas demais.

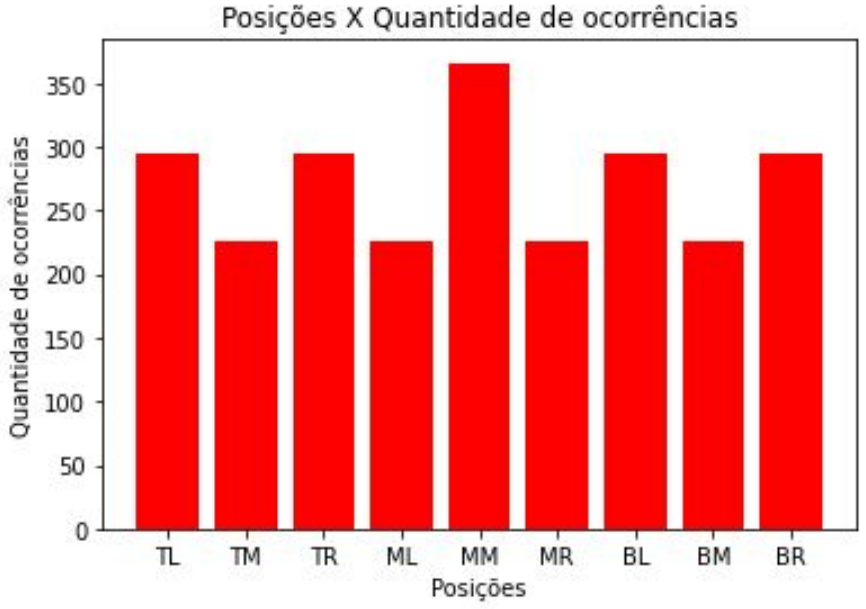

Figura 10: Gráfico que informa a quantidade de ocorrências por posição.
Fonte: Própria.

Com o intuito de desenvolver o experimento, foi feita a divisão da base em 3 partes iguais. A primeira parte, se refere aos dados para o treinamento, a segunda dados de predição, a terceira validação. Como a base de dados estava dividida ao meio pela atributo de classificação, foi necessário obter a quantidade de ocorrências de cada uma delas individualmente e somar. Então, cada parte das três divisões, está composta por aproximadamente 208 capacidades de vitória, e 110 de derrota, totalizando cerca de 318 valores por parte.

O treinamento foi definido com 250 épocas e uma taxa de aprendizado de 0,12. A rede no treinamento obteve uma taxa de acerto médio de 99,15% e um erro médio de 0,02. Quando submetida a outro conjunto, seu acerto médio foi reduzido para 69,91% e erro médio de 5,82. Conforme a imagem abaixo, é possível verificar como o acerto e erro médio se comportam ao longo das épocas.

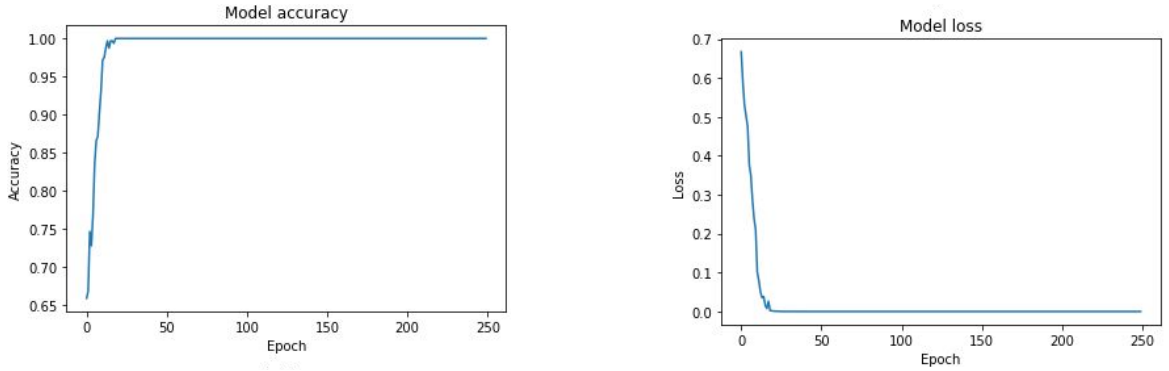

Figura 11: Acerto médio e erro médio ao decorrer das épocas.
Fonte: Própria.

Assim como no experimento anterior, foi utilizado perceptron de multicamada (MLP), com 3 camadas escondidas (12, 4 e 1 nodes) sendo a última de saída. A função de ativação sigmóide foi aplicada em todas as camadas, bem como o ‘RMSprop’’ como otimizador da rede. Como parâmetro de perda, foi utilizado o ‘binary_crossentropy’. Os hiperparâmetros citados foram escolhidos depois da rede apresentar os melhores resultados no treinamento.

O treinamento apresentou bons resultados. Entretanto, quando submetida a outro conjunto de dados, a rede obteve queda significativa na sua capacidade assertiva. Sendo assim, a mesma possui uma relativa capacidade de identificar se é possível o jogador “X” ganhar o jogo, mas não se deve confiar 100%.

Para que fosse possível o estudo dos dados foram utilizados: Adaptive Moment Estimation (Adam), função Sigmoid, binary_crossentropy e RMSprop.

Adam trata-se de um método que calcula as taxas de aprendizado adaptáveis para cada

parâmetro. Já a função sigmoid foi utilizada para organizar parâmetros que assumiam 0 ou 1, diferente do binary_crossentropy que utilizado em decisões de sim ou não, e por último o RMSpropé que tem como ideia central manter a média móvel dos gradientes quadrados para cada peso e então dividimos o gradiente por raiz quadrada do quadrado médio. É por isso que se chama RMSprop (root mean square).