Modelagem de Redes Neurais para Previsão de Custos em Seguros de Saúde

Gabriel D'assumpção de Carvalho

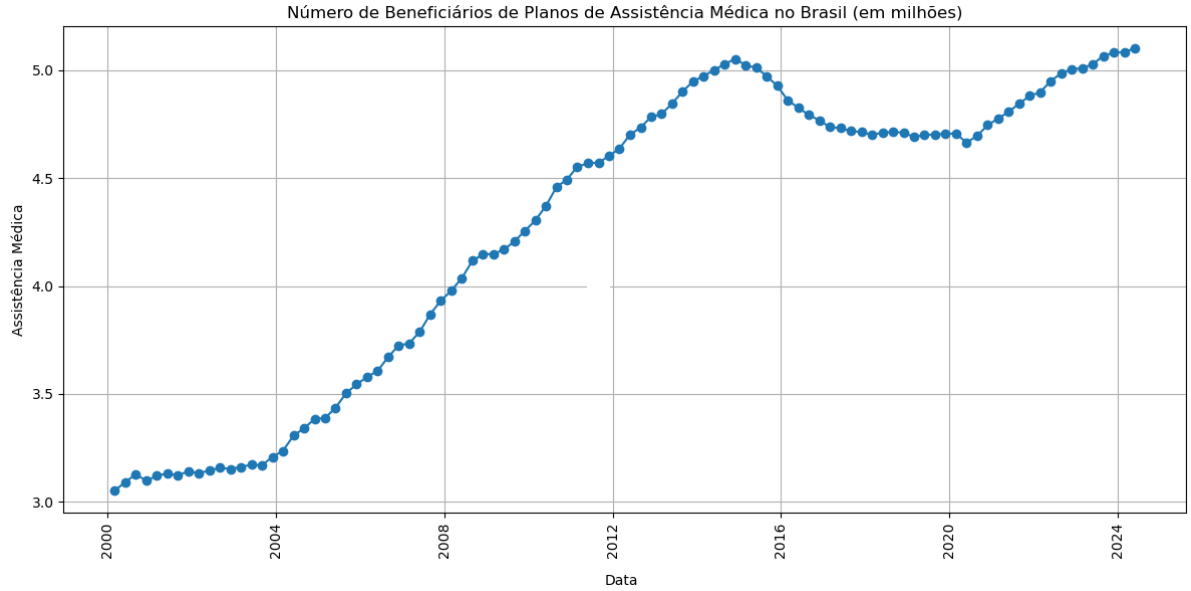
19/08/2024

1 Introdução

As doenças cardiovasculares continuam sendo a principal causa de mortalidade no Brasil, de acordo com o estudo Global Burden of Disease (GBD) de 2019 e os dados do Sistema Único de Saúde (SUS). O preocupante aumento na incidência de ataques cardíacos entre jovens e adultos é amplamente atribuído à falta de atividade física e a uma dieta inadequada. Esse panorama destaca a urgência de promover hábitos de vida mais saudáveis, seguindo a recomendação de Michael Pollan: "descasque mais, desembale menos", que incentiva o consumo de alimentos frescos e não processados.

A deterioração da saúde pública tem impulsionado o crescimento do setor de planos de saúde. De acordo com a Agência Nacional de Saúde Suplementar (ANS), o Brasil registrou um aumento de aproximadamente 20,5 milhões de novos beneficiários de planos de assistência médica entre março de 2000 e junho de 2024.

Figura 1: Gráfico mostrando o crescimento dos beneficiários de planos de saúde no Brasil



Fonte: Agência Nacional de Seguro (ANS)

O gráfico acima ilustra o aumento no número de beneficiários ao longo de 24 anos, evidenciando a crescente demanda por cuidados médicos. Esse crescimento torna essencial a análise e previsão das despesas associadas aos seguros de saúde, objetivo principal deste projeto.

Para realizar esta análise, será utilizada uma base de dados disponível no **Kaggle**. A base contém 1.338 observações e 6 características, com o custo do seguro médico como variável alvo. A seguir, estão as variáveis da base de dados:

Atributo	Descrição	Tipo
Idade	A idade da pessoa segurada.	Numérica - Inteiro
Sexo	Gênero (masculino ou feminino) do segurado.	Categórico - Binário
IMC	Índice de Massa Corporal: uma medida de gordura corporal baseada na altura e no peso.	Numérica - Contínuo
Crianças	O número de dependentes cobertos.	Numérica - Discreta
Fumante	Se o segurado é fumante (sim ou não).	Categórico - Binário
Região	A área geográfica de cobertura.	Categórico - Nominal
Encargos	Os custos do seguro médico incorridos pelo segurado.	Numérica - Contínuo

Tabela 1: Descrição das variáveis da base de dados

2 Fundamentos Teóricos e Metodológicos

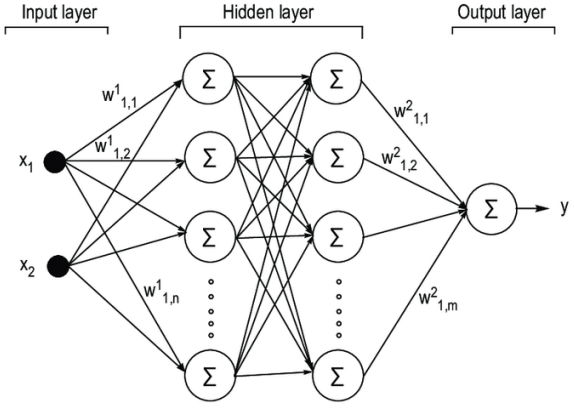
O objetivo deste trabalho é prever os custos do seguro médico com base nas características dos clientes, configurando-se como um problema de regressão. A abordagem tradicional de regressão linear pode ser representada pela seguinte equação:

$$\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \epsilon \tag{1}$$

Entretanto, as suposições subjacentes à regressão linear, como a normalidade das variáveis, a independência entre as características dos clientes e a normalidade dos resíduos, podem não ser atendidas neste contexto. Por essa razão, este estudo opta por utilizar redes neurais, especificamente o Perceptron Multicamadas (MLP), para realizar a previsão dos custos do seguro.

Modelos com múltiplas camadas, como o MLP (Perceptron Multicamadas), são mais adequados para resolver problemas não lineares e de alta complexidade. Redes com $S \in \mathbb{N}$ neurônios são compostas por camadas de neurônios, onde cada camada pode conter um conjunto de neurônios $O = \{o_1, o_2, \ldots, o_S\}$, vieses $B = \{b_1, b_2, \ldots, b_S\}$ e pesos $W = \{w_{i,j}\}$, que conectam os neurônios entre as camadas. Os pesos $w_{i,j}$ representam a força das conexões entre o neurônio i de uma camada e o neurônio j da próxima camada. O MLP é estruturado em camadas: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio em uma camada é ativado por uma função de ativação que pode ser linear ou não linear. A Figura 2 ilustra a arquitetura do MLP.

Figura 2: Arquitetura do Perceptron Multicamadas



Fonte: `https://www.researchgate.net/figure/Schematic-structure-of-a-perceptron-neural-network_fig2_260291992`

A fórmula geral para o cálculo da ativação de um neurônio j em uma camada oculta é:

$$\phi_j(\text{net}_j) = \phi_j \left(\sum_{k=1}^{n} w_{k,j} \cdot o_k + b_j \right) \tag{2}$$

onde:

$$\text{net}_j = \sum_{k=1}^{n} w_{k,j} \cdot o_k + b_j \tag{3}$$

Aqui, $w_{k,j}$ são os pesos entre o neurônio k da camada anterior e o neurônio j da camada atual, o_k são as saídas dos neurônios da camada anterior e b_j é o viés do neurônio j.

O modelo MLP é capaz de capturar interações complexas entre as variáveis, tornando-o mais eficaz do que a regressão linear tradicional para dados que não seguem padrões simples.

2.1 Análise Exploratória de Dados

Esta seção apresenta uma análise exploratória dos dados com o objetivo de identificar relações e padrões entre as variáveis explicativas e a variável resposta, *charges* (custos do seguro saúde). A análise se divide em duas etapas principais:

2.1.1 Relações entre Variáveis Explicativas e Variável Resposta

Inicialmente, investigamos as relações entre as variáveis explicativas e a variável resposta utilizando testes estatísticos apropriados.

Variáveis Categóricas Binárias (Sexo e Fumante): Para avaliar a relação entre as variáveis categóricas binárias (*sex* e *smoke*) e as variáveis numéricas (idade, IMC) e a variável resposta (*charges*), utilizamos o teste t de Student. Os resultados são apresentados na Tabela 2.

Tabela 2: Resultados do Teste t de Student para Variáveis Categóricas Binárias

Variáveis	**Estatística t**	**Valor Crítico t (0.05)**	**Conclusão**
Idade - Sexo	-0,7625	1,9617	Não significativo
Idade - Fumante	-0,9210	1,9617	Não significativo
IMC - Sexo	1,6970	1,9617	Não significativo
IMC - Fumante	0,1335	1,9617	Não significativo
Charges - Sexo	**2,1001**	1,9617	**Significativo**
Charges - Fumante	**32,7519**	1,9617	**Significativo**

Os resultados do teste t, apresentados na Tabela 2, demonstram que:

- **Não há diferença estatisticamente significativa** na idade e no IMC entre os diferentes sexos e entre fumantes e não fumantes.
- **Há uma diferença estatisticamente significativa** nos custos do seguro (*charges*):
  - Homens apresentam custos de seguro significativamente maiores em comparação às mulheres.
  - Fumantes apresentam custos de seguro significativamente maiores em comparação aos não fumantes.

Variável Categórica Multiclasse (Região): Para analisar a relação entre a variável categórica multiclasse *region* (sudeste, sudoeste, noroeste, nordeste) e as variáveis numéricas (idade, IMC, número de filhos) e a variável resposta (*charges*), empregamos o teste ANOVA. A Tabela 3 exibe os resultados.

Tabela 3: Resultados do Teste ANOVA para a Variável Região

Variáveis	**Estatística F**	**Valor Crítico F (0.05)**	**Conclusão**
Idade - Região	0,0798	2,6116	Não significativo
IMC - Região	**39,4951**	2,6116	**Significativo**
Filhos - Região	0,7175	2,6116	Não significativo
Charges - Região	**2,9696**	2,6116	**Significativo**

Os resultados do teste ANOVA, apresentados na Tabela 3, indicam que:

- **A região de residência não influencia significativamente** a idade e o número de filhos.
- **O IMC varia significativamente entre as diferentes regiões**. Observa-se um aumento do IMC nas regiões Sul, o que pode estar relacionado a fatores socioeconômicos.
- **Os custos do seguro também variam significativamente entre as regiões**, com a região Sul apresentando custos mais elevados.

Variáveis Numéricas (Idade e IMC): Para investigar a relação entre as variáveis numéricas (idade, IMC e filhos) e a variável resposta (*charges*), utilizamos a correlação de Pearson para avaliar a relação linear e a correlação de Spearman para avaliar a relação monotônica Tabelas 4.

Tabela 4: Correlações entre Variáveis Numéricas e Custos do Seguro

Variáveis	**Correlação de Pearson**	**Correlação de Spearman**
Charges - Idade	0,2990	**0,5344**
Charges - IMC	0,1983	0,1194
Charges - Filhos	-	0,1333

As análises de correlação, apresentadas na Tabela 4, revelam que:

- **Tanto a idade quanto o IMC apresentam uma correlação positiva com os custos do seguro**, indicando que os custos tendem a aumentar com o aumento da idade e do IMC.
- **A correlação com a idade é mais forte do que com o IMC**, sugerindo que a idade é um preditor mais relevante para os custos do seguro.
- **A correlação de Spearman entre idade e custos do seguro (0,5344) é mais forte do que a correlação de Pearson (0,2990)**, sugerindo uma relação não linear entre essas variáveis.

2.1.2 Discussão

A análise exploratória dos dados revelou insights importantes sobre os fatores que influenciam os custos do seguro saúde. Observamos que:

- Fatores demográficos como sexo, idade e região de residência, assim como hábitos de saúde como o tabagismo, estão relacionados aos custos do seguro.
- A idade é um preditor mais relevante para os custos do seguro do que o IMC.
- As regiões Sul apresentam custos de seguro mais elevados, o que pode estar associado a fatores socioeconômicos e disparidades no acesso à saúde.

É crucial considerar esses insights durante o desenvolvimento do modelo de redes neurais para garantir que ele seja capaz de capturar e modelar adequadamente as relações complexas presentes nos dados.

3 Tratamento dos Dados

Após a análise exploratória, procedemos ao tratamento dos dados. A ausência de valores ausentes ou duplicados no conjunto de dados direcionou o tratamento para a transformação das variáveis categóricas em variáveis dummy, resultando em um total de 11 variáveis explicativas. Em seguida, as variáveis ***age***, ***imc*** e ***charges*** foram normalizadas para o intervalo de 0 a 1. Os dados foram então divididos em conjuntos de treinamento (80%) e teste (20%).

4 Arquitetura da Rede Neural

Foi desenvolvido um modelo de rede neural totalmente conectada (Fully Connected Network) com a seguinte estrutura:

- **Camada de Entrada:** 11 neurônios, representando as 11 variáveis explicativas após o pré-processamento dos dados.
- **Camadas Ocultas:** 4 camadas densas, com as seguintes características:
  - Camada 1: 128 neurônios, função de ativação ReLU, seguida por uma camada de Dropout com taxa de 0.1.
  - Camada 2: 256 neurônios, função de ativação ReLU, seguida por uma camada de Dropout com taxa de 0.1.

  - Camada 3: 800 neurônios, função de ativação ReLU, seguida por uma camada de Dropout com taxa de 0.1.
  - Camada 4: 256 neurônios, função de ativação ReLU.
- **Camada de Saída:** 1 neurônio, função de ativação linear, para predição da variável contínua *charges*.

A escolha da função de ativação ReLU para as camadas ocultas visa introduzir não linearidade ao modelo, enquanto a função linear na camada de saída permite a predição de valores contínuos.

5 Treinamento e Validação do Modelo

O treinamento da rede neural foi realizado utilizando o otimizador Adam, que ajusta os pesos da rede de forma iterativa com o objetivo de minimizar a função de perda Erro Quadrático Médio (MSE). O modelo foi treinado por 10.000 épocas, utilizando 20% dos dados de treinamento para validação a cada época. Essa validação durante o treinamento permite monitorar o desempenho do modelo em dados não vistos e ajustar os hiperparâmetros para evitar overfitting, buscando um bom equilíbrio entre o erro nos dados de treinamento e nos dados de validação.

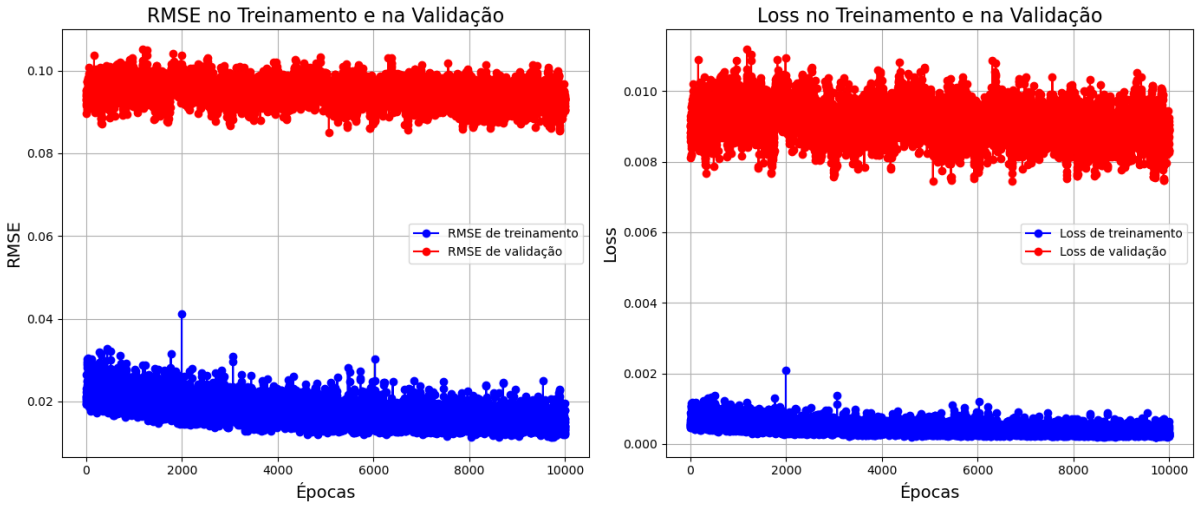
6 Avaliação do Modelo

A avaliação do modelo de regressão foi realizada considerando as métricas RMSE (Root Mean Squared Error) e MSE (Mean Squared Error). Além da análise quantitativa, foi conduzida uma análise visual do desempenho do modelo por meio de um gráfico de dispersão, comparando os valores preditos com os valores reais.

6.1 Desempenho durante o Treinamento

A Figura 3 ilustra a evolução do RMSE e da função de perda (MSE) ao longo das épocas de treinamento, tanto para os dados de treinamento quanto para os dados de validação.

Figura 3: RMSE e Função de Perda (MSE) durante o Treinamento e a Validação



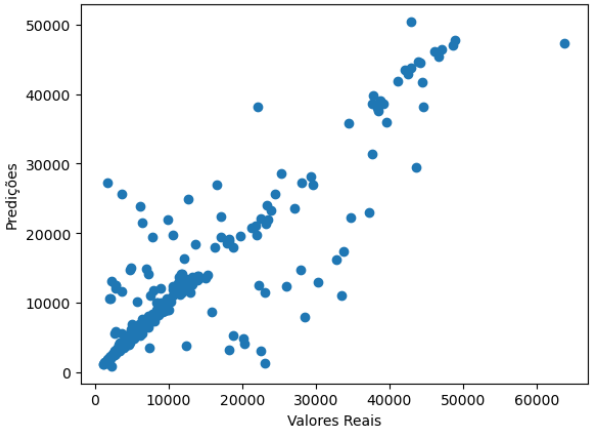
- **Análise dos Resultados do Treinamento:**
  - **Diferença entre Treinamento e Validação:** Observa-se que as curvas de RMSE e MSE para os dados de validação estão, como esperado, acima das curvas correspondentes aos dados de treinamento. Essa diferença, embora presente, não é muito acentuada, sugerindo que o modelo generaliza razoavelmente bem para dados não vistos, sem apresentar overfitting severo.

– **Proximidade de Zero:** A proximidade dos valores de RMSE e MSE a zero é um reflexo direto da normalização da variável resposta para o intervalo de 0 a 1, aplicada durante o pré-processamento dos dados.

6.2 Capacidade Preditiva do Modelo

A Figura 4 apresenta o gráfico de dispersão dos valores preditos versus os valores reais para os dados de teste, fornecendo uma visualização da capacidade preditiva do modelo.

Figura 4: Gráfico de Dispersão dos Valores Preditos e Reais dos Dados de Teste



- **Análise da Capacidade Preditiva:**
  - **Tendência Linear com Dispersão:** O gráfico de dispersão exibe uma tendência linear, indicando que o modelo, em geral, realiza predições coerentes com os valores reais. No entanto, observa-se uma dispersão dos pontos em torno da reta ideal (onde a predição seria igual ao valor real), sugerindo que o modelo apresenta erros em suas previsões. A análise da dispersão pode auxiliar na identificação de regiões onde o modelo apresenta maior dificuldade em predizer os valores corretamente.

7 Conclusão

Este estudo demonstrou a eficácia das redes neurais artificiais, particularmente o modelo Perceptron Multicamadas (MLP), na previsão de custos de seguros de saúde. A análise dos dados evidenciou a influência de fatores como idade, sexo, tabagismo e região, destacando a complexidade do problema.

O MLP, com sua estrutura de múltiplas camadas e funções de ativação não lineares, mostrou-se capaz de modelar essas relações complexas, superando as limitações da regressão linear. O uso do otimizador Adam e da regularização com dropout contribuiu para um treinamento eficiente e um modelo com boa generalização.

A avaliação do modelo, com RMSE e MSE, indicou um bom desempenho, embora haja margem para melhorias. As descobertas têm potencial para aprimorar a gestão de riscos, personalizar planos de saúde e apoiar políticas públicas de promoção da saúde. A continuidade da pesquisa poderá fortalecer ainda mais os modelos preditivos no setor de saúde.

8 Referências

- AL-MA'AMARI, M. Deep neural networks for regression problems. Disponível em: `https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33`. Acesso em: 25 set. 2024.
- OLIVEIRA, G. M. M. DE et al. Estatística cardiovascular – brasil 2021. Arquivos brasileiros de cardiologia, v. 118, n. 1, p. 115–373, 2022.