Redes Neurais para Previsão de Custos em Seguros de Saúde

Gabriel D'Assumpção de Carvalho
gabriel.dassumpcao@ufpe.br

Ciências Atuarias-UFPE

30 de setembro de 2024

Sumário

Introdução

Introdução ao Modelo Perceptron

Doenças cardiovasculares são a principal causa de morte no Brasil.

Aumento de 20,5 milhões de beneficiários de planos de saúde entre 2000 e 2024 (ANS).

Previsão de custos de seguros de saúde é crucial.

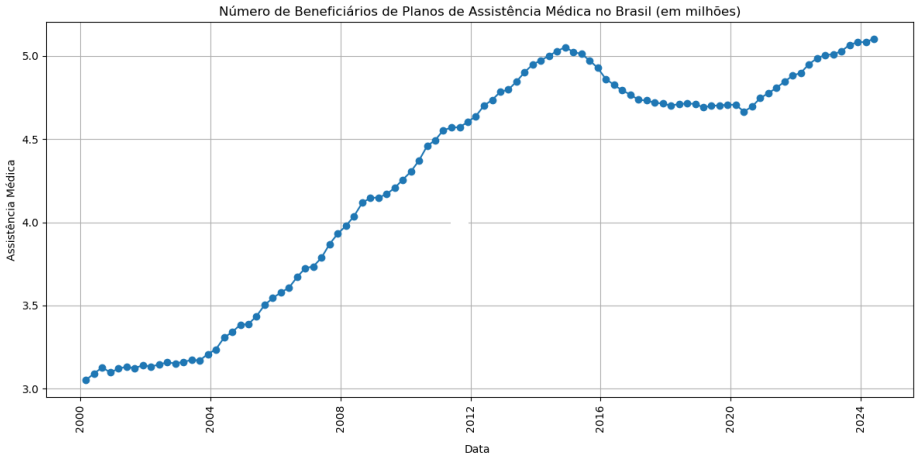


Figura: Fonte: Agência Nacional de Saúde (ANS)

Análise Exploratória

Base de Dados

Fonte: Kaggle

1.338 observações, 6 características.

Variável alvo: custos de seguro (*charges*).

Atributo	**Descrição**	**Tipo**	**Dados Nulos**
Age	A idade da pessoa segurada.	Numérica - Inteiro	Não
Sex	Gênero (masculino ou feminino) do segurado.	Categórico - Binário	Não
IMC	Índice de Massa Corporal: uma medida de gordura corporal baseada na altura e no peso.	Numérica - Contínuo	Não
Children	O número de dependentes cobertos.	Numérica - Discreta	Não
Smoke	Se o segurado é fumante (sim ou não).	Categórico - Binário	Não
Region	A área geográfica de cobertura.	Categórico - Nominal	Não
Charges	Os custos do seguro médico incorridos pelo segurado.	Numérica - Contínuo	Não

Tabela: Descrição das variáveis da base de dados

Estatísticas Descritivas da Base de Dados

Tabela: Dados Duplicados

age	sex	imc	children	smoke	region	charges
19	male	30,59	0	no	northwest	1.639,56

Tabela: Descrição de Variáveis Numéricas

	age	imc	children	charges
count	1.338	1.338	1.338	1.338
mean	39,21	30,66	1,09	13.270,42
std	14,05	6,1	1,21	12.110,01
min	18	15,96	0	1.121,87
25%	27	26,3	0	4.740,29
50%	39	30,4	1	9.382,03
75%	51	34,69	2	16.639,91
max	64	53,13	5	63.770,43

Estatísticas Descritivas da Base de Dados

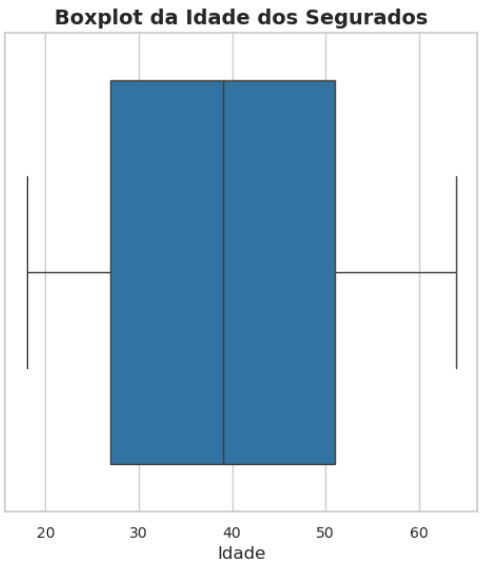
Tabela: Modas

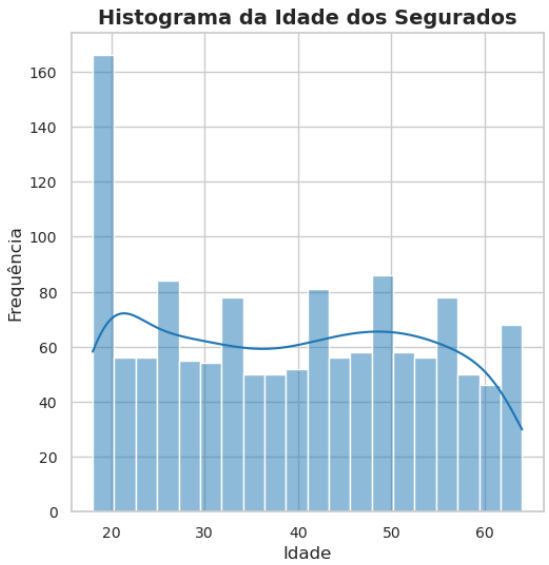
	age	imc	children	charges
Moda	18	32,3	0	1.639,56

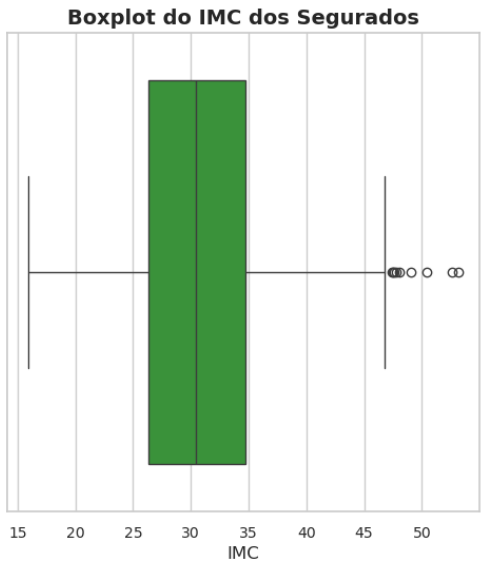
Tabela: Descrição de Variáveis Categóricas

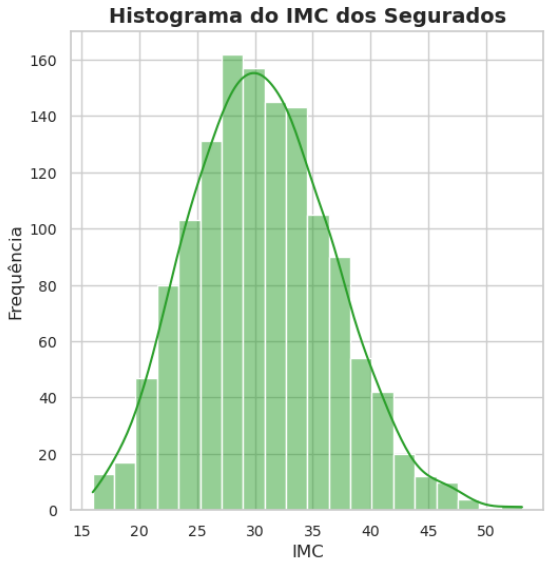
	sex	smoke	region
count	1.338	1.338	1.338
unique	2	2	4
top	male	no	southeast
freq	676	1.064	364

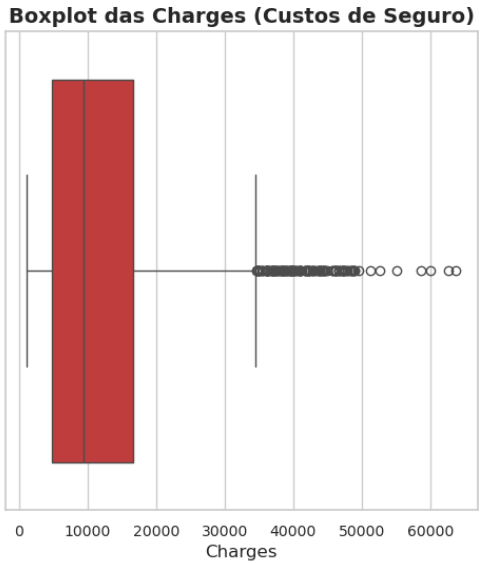
Gráfico Variáveis Numéricas











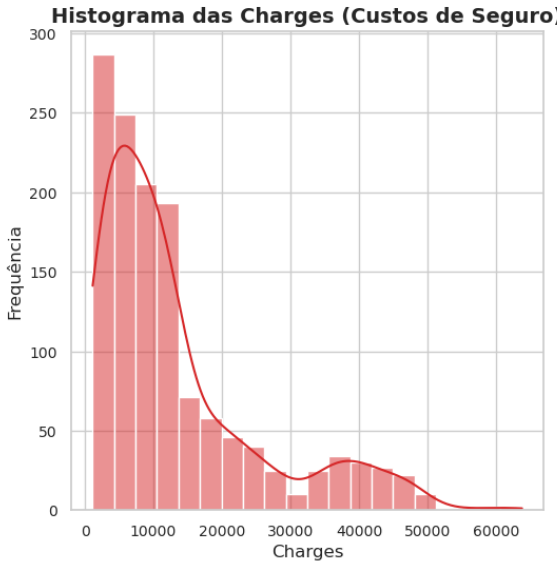
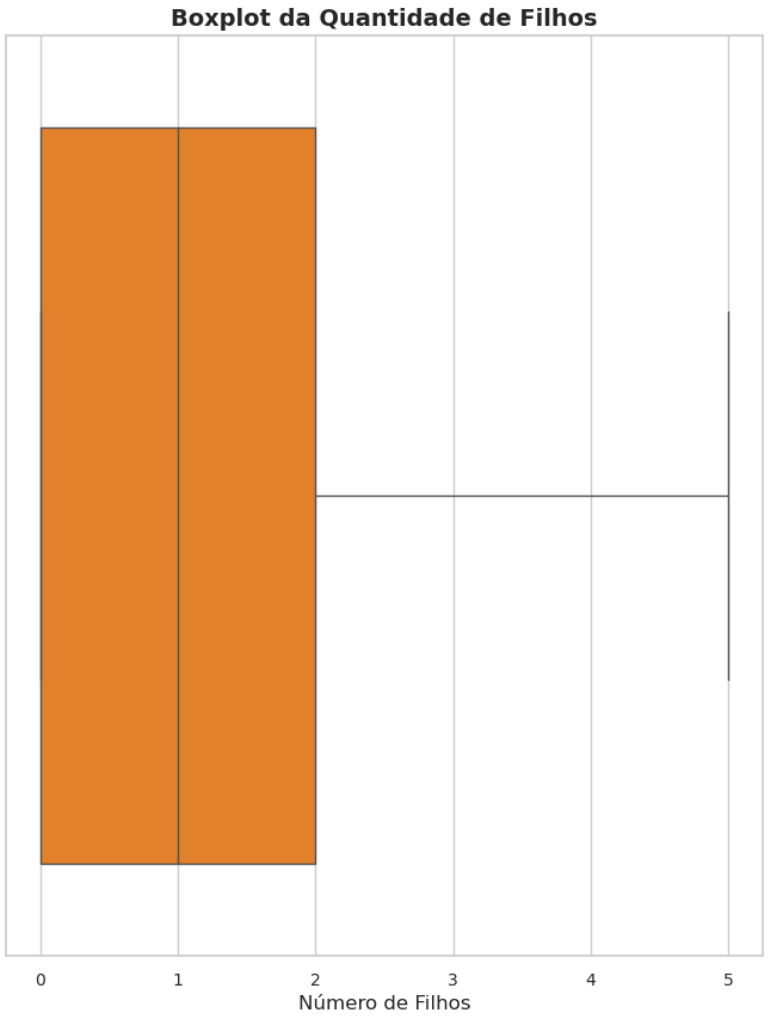


Gráfico Variável Numérica



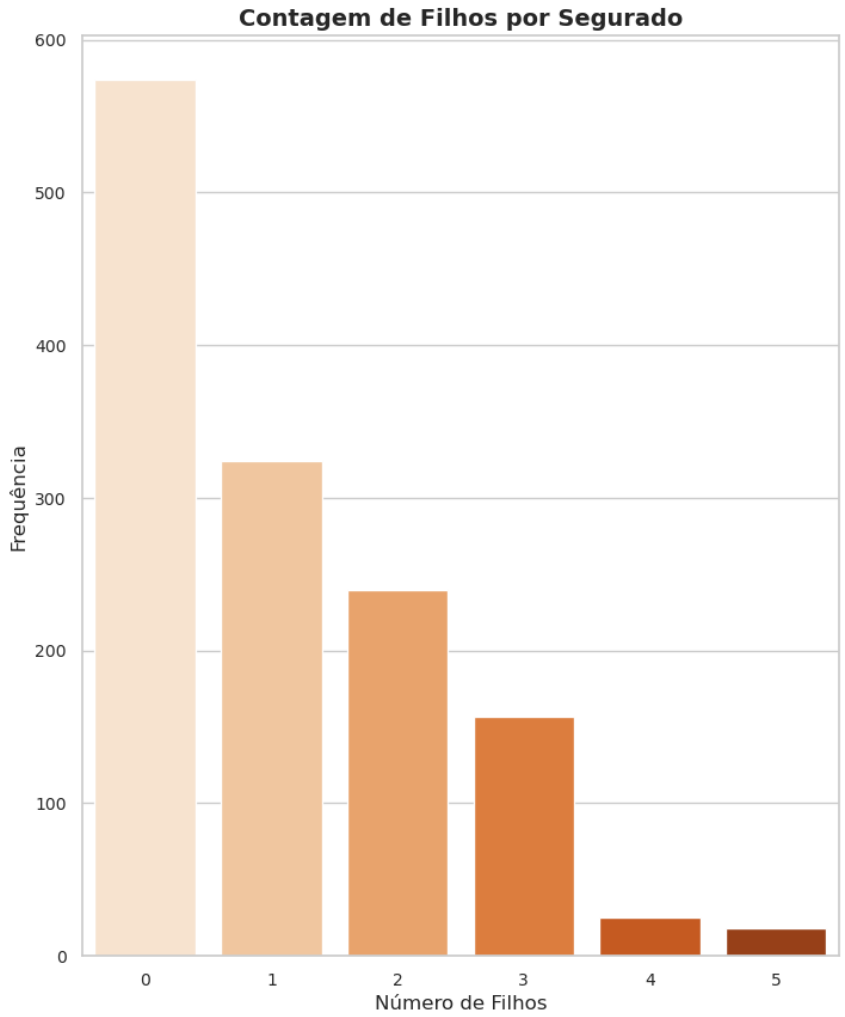
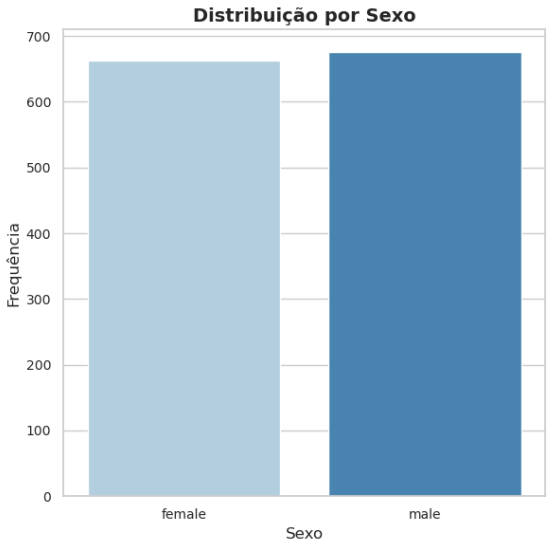
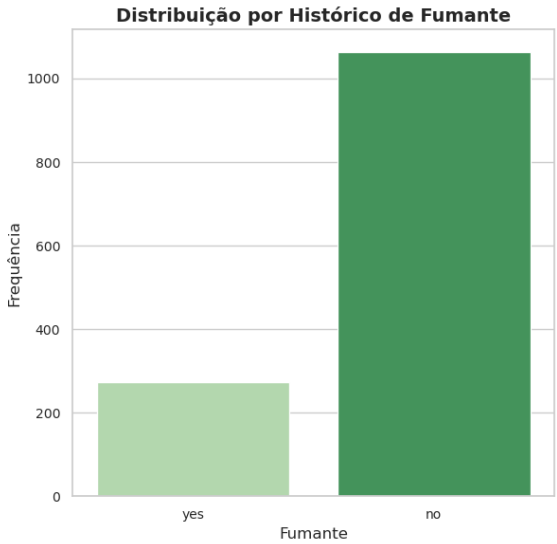
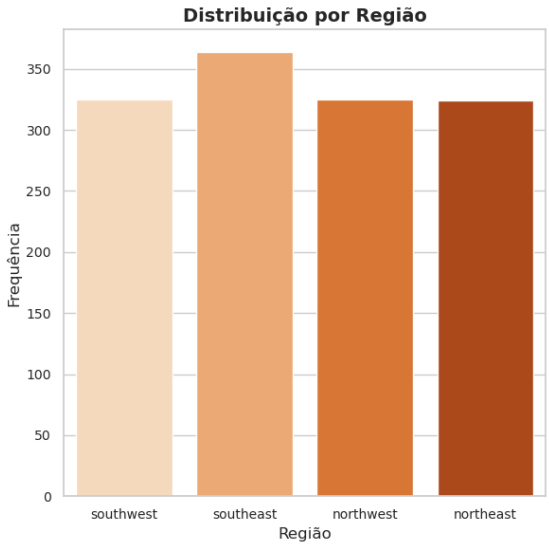


Gráfico Variáveis Categóricas







Resultados do Teste t de Student para Variáveis Categóricas Binárias

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

Variáveis	**Estatística t**	**Valor Crítico t (0.05)**	**Conclusão**
Age - Sex	-0,76	1,96	Não significativo
Age - Smoke	-0,92	1,96	Não significativo
IMC - Sex	1,70	1,96	Não significativo
IMC - Smoke	0,13	1,96	Não significativo
Charges - Sexo	**2,10**	1,96	**Significativo**
Charges - Smoke	**32,75**	1,96	**Significativo**

Resultados do Teste ANOVA para a Variável Região

Tabela: Resultados do Teste ANOVA para a Variável Região

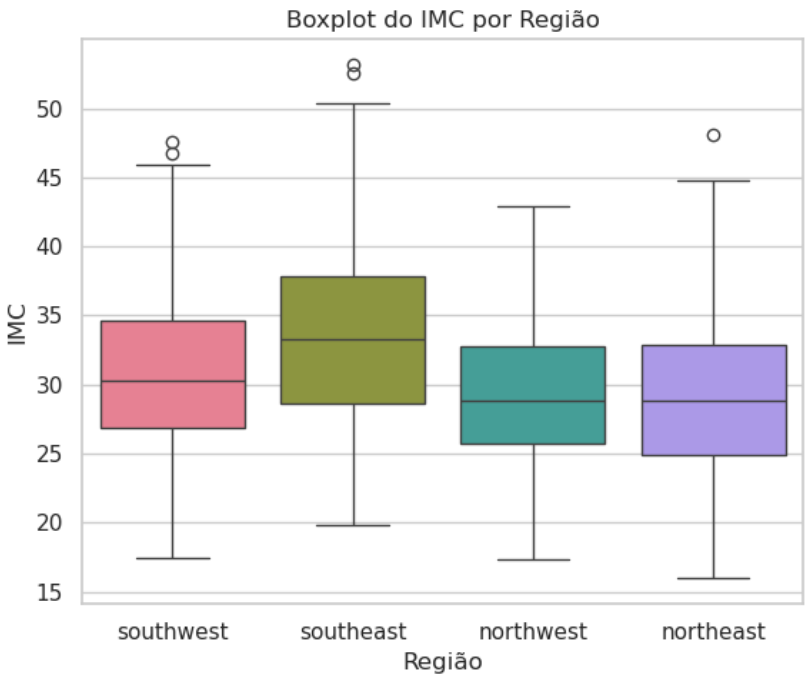
Variáveis	**Estatística F**	**Valor Crítico F (0,05)**	**Conclusão**
Age - Region	0,08	2,61	Não significativo
IMC - Region	**39,50**	2,61	**Significativo**
Children - Region	0,72	2,61	Não significativo
Charges - Region	**2,97**	2,61	**Significativo**

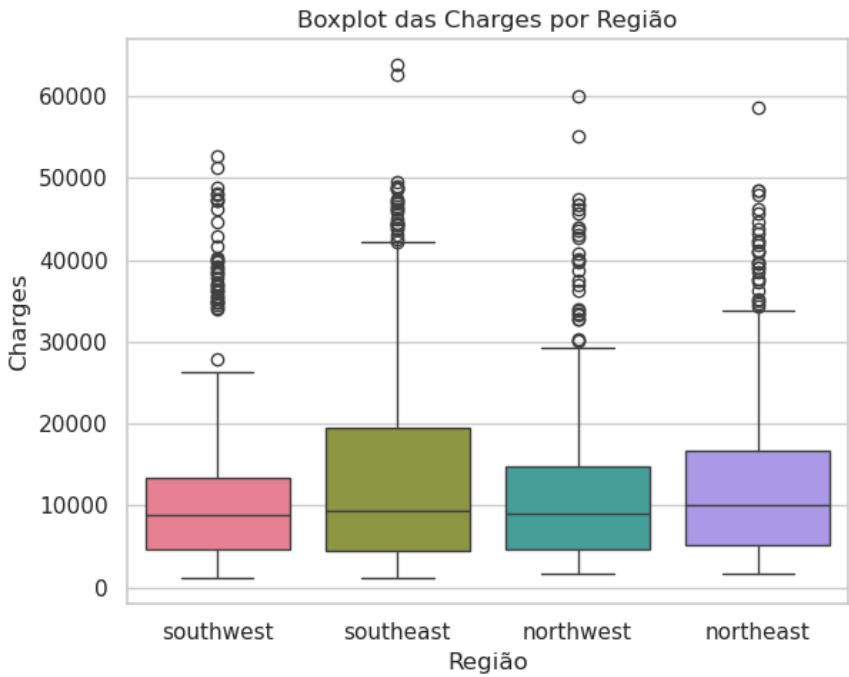
Correlações entre Variáveis Numéricas e Custos do Seguro

Tabela: Correlações entre Variáveis Numéricas e Custos do Seguro

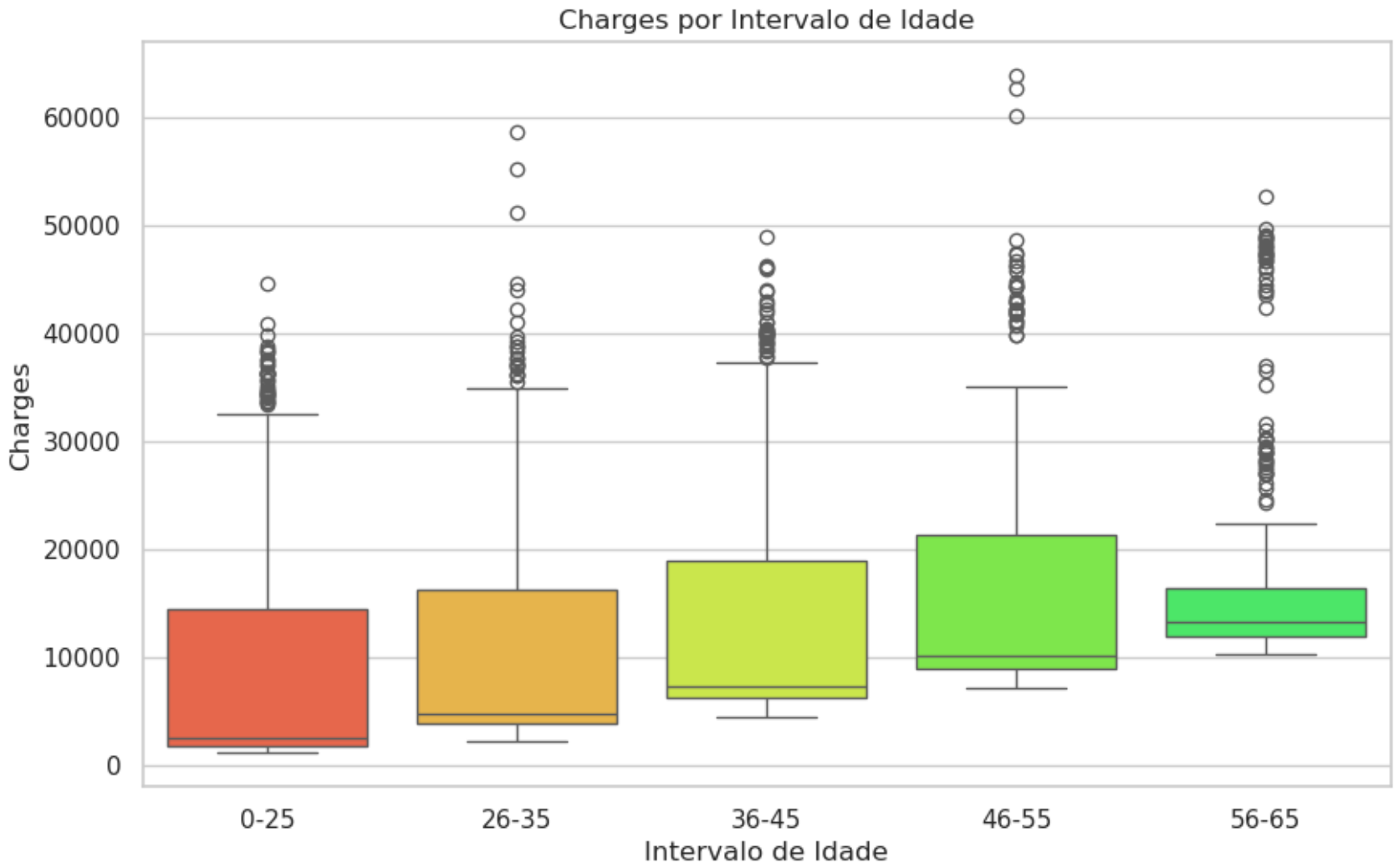
Variáveis	**Correlação de Pearson**	**Correlação de Spearman**
Charges - Age	0,30	**0,53**
Charges - IMC	0,20	0,12
Charges - Children	-	0,13

Boxplots de IMC e Charges por Região





Boxplots Charges por Faixa Etária



Modelo da Rede Neural

Treinamento e Validação do Modelo

Treinamento da Rede Neural:

80% Treinamento e 20% Teste.
Utilização do otimizador **Adam**.
Função de perda: **Erro Quadrático Médio (MSE)**.
10.000 épocas de treinamento.
20% dos dados de treinamento para validação em cada época.

Tabela: Estrutura do Modelo Keras

Camada	**Tipo**	**Ativação**	**Saída**	**Parâmetros**
Densa 1	Dense (128)	ReLU	(None, 128)	12.032
Dropout 1	Dropout (10%)	-	(None, 128)	0
Densa 2	Dense (256)	ReLU	(None, 256)	33.024
Dropout 2	Dropout (10%)	-	(None, 256)	0
Densa 3	Dense (800)	ReLU	(None, 800)	205.600
Dropout 3	Dropout (10%)	-	(None, 800)	0
Densa 4	Dense (256)	ReLU	(None, 256)	204.856
Densa 5	Dense (1)	Linear	(None, 1)	257
Total				**455.769**

Desempenho durante o Treinamento

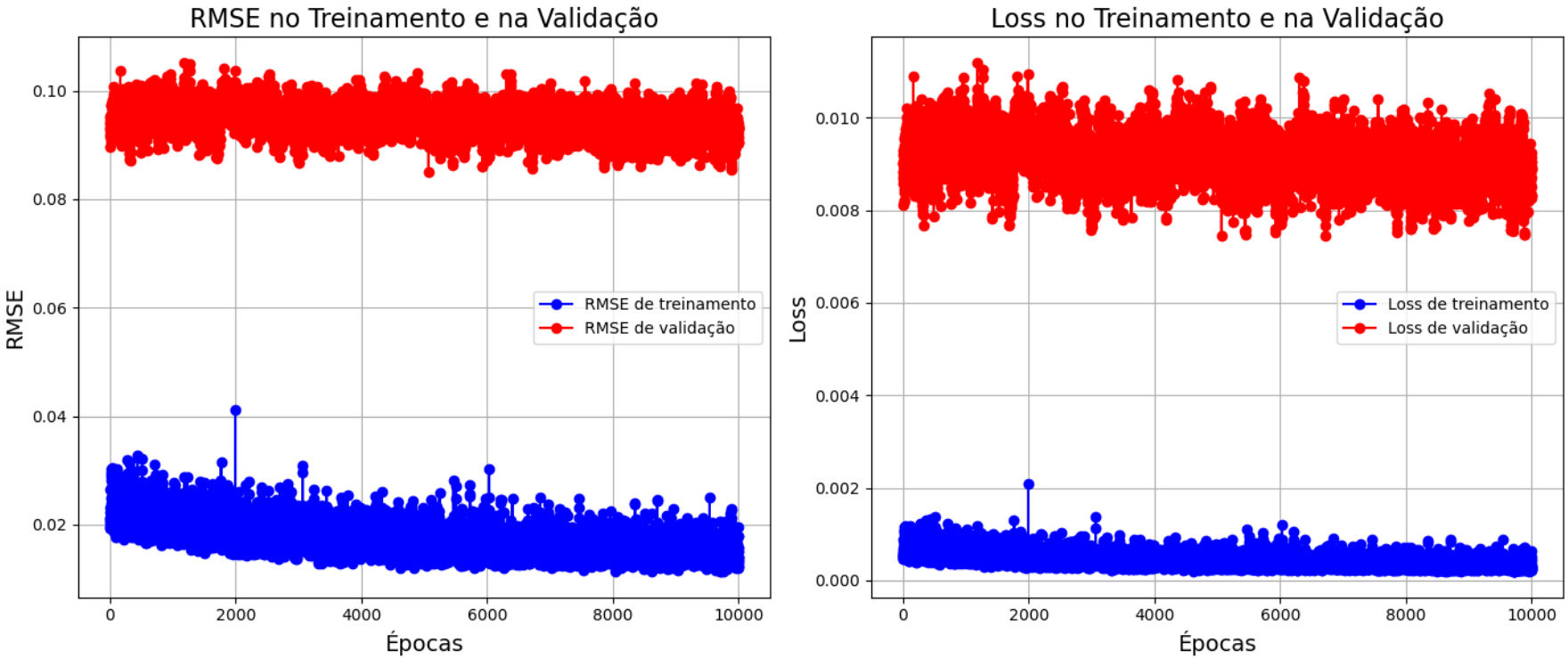


Figura: Evolução do RMSE e da função de perda (MSE) ao longo das épocas.

Capacidade Preditiva do Modelo

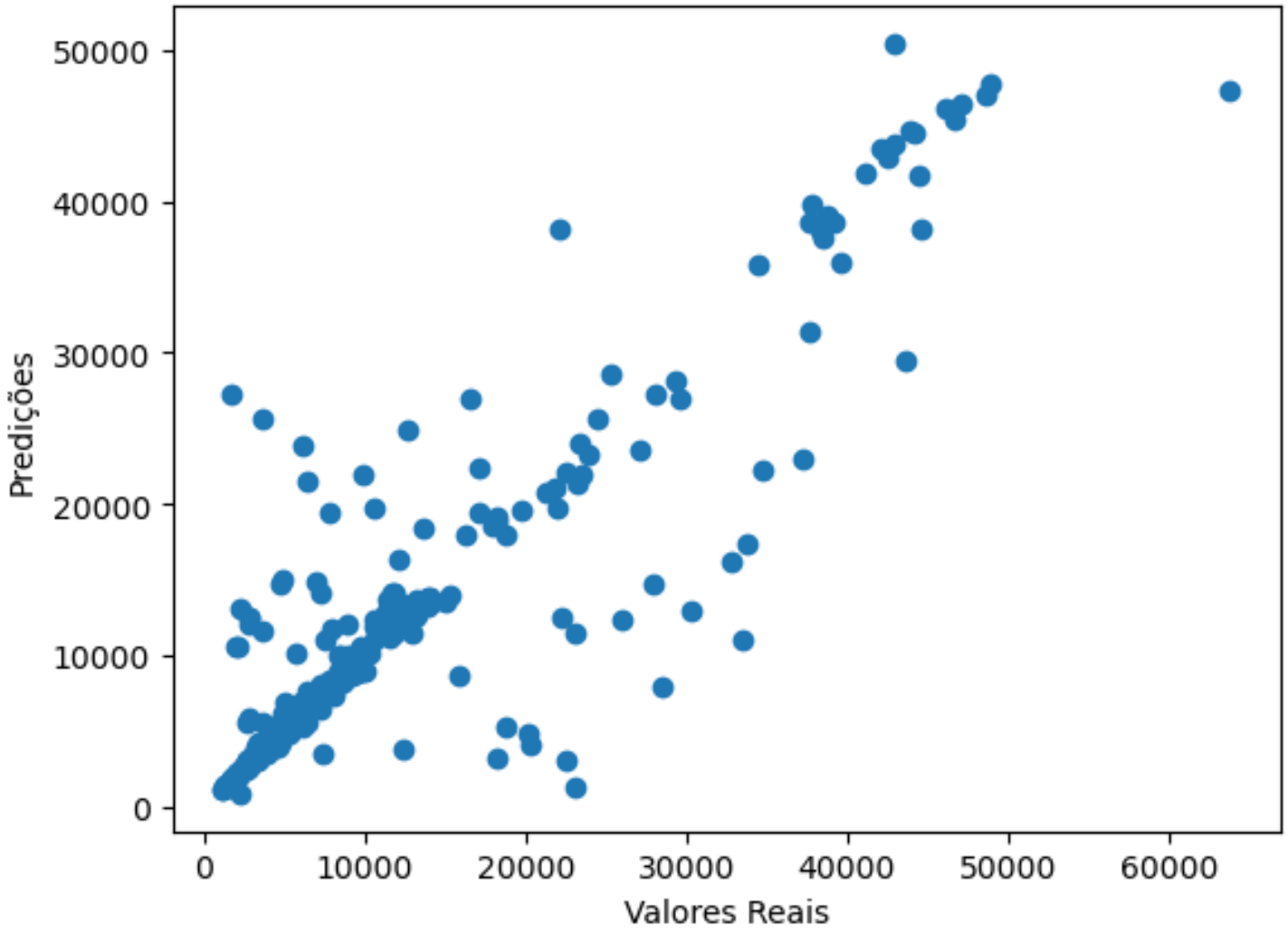


Figura: Gráfico de dispersão dos valores preditos versus reais para dados de teste.

Conclusão

Considerações Finais

Eficácia das Redes Neurais:

O modelo Perceptron Multicamadas (MLP) mostrou-se eficaz na previsão dos custos de seguros de saúde.
Fatores como **idade, sexo, tabagismo e região** influenciam significativamente os resultados.

Avaliação do Desempenho:

Os índices de RMSE e MSE sugerem um desempenho satisfatório do modelo, embora ainda haja oportunidades para melhorias.

Sugestões de Melhoria:

Implementação de um **comitê heterogêneo de redes neurais** para aprimorar a precisão das previsões.
Exploração de combinações matemáticas entre variáveis correlacionadas para potencializar os resultados.
Remoção de outliers nas variáveis **idade, sexo, tabagismo e região** para melhorar a qualidade dos dados.

Potencial das Descobertas:

As descobertas podem contribuir significativamente para a gestão de riscos e a personalização de planos de saúde.

Referências

AL-MA'AMARI, M. Deep neural networks for regression problems. Disponível em: https://towardsdatascience.com/deep-neural-networks-for-regression 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.