Projeto Final: Análise de dados de Plano de Saúde em R

UNIT - Universidade Tiradentes

Curso: Ciência de Dados e Inteligência Artificial

Disciplina: Linguagens Python e R com Foco em Análise de Dados

2025

Fernanda Amaral de Souza; Larissa Castor Ramos

Prof. Msc. Bruna Martini Dalmoro

Este relatório investiga determinantes de preço e de utilização de planos de saúde a partir de uma base padronizada (n = 1.338), combinando etapas descritivas e modelos estatísticos.

Foi verificado, inicialmente, os clientes que pagam o menor e o maior preço, respectivamente, no plano de saúde. Servindo como ponto de referência mínimo e máximo com relação ao preço para o banco de dados.

Tabela 1: Identificação das características dos clientes que pagam o menor e o maior preço do plano de saúde.

	idade	**sexo**	**imc**	**dependentes**	**tabagismo**	**regiao**	**preco**	**uso**
Menor preço	18	masculino	23,21	0	0	sudeste	1121,874	0
Maior preço	39	feminino	18,3	5	1	sudoeste	23859,14	1

Ao avaliar a quantidade total de cliente do sexo masculino e feminino, foram identificados 676 registros do sexo masculino e 662 registros do sexo feminino no conjunto de dados.

Foi realizada uma análise de regressão linear múltipla com o objetivo de identificar e quantificar o impacto das variáveis sobre o preço dos planos de saúde. Na Tabela 2 estão apresentados os parâmetros estipulados pela regressão linear.

Tabela 2: Coeficientes estimados do modelo de regressão linear múltipla para o preço dos planos de saúde

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	-6957,33	988,4643	-7,03853	3,1E-12
Idade	256,8955	11,90662	21,57586	9,25E-89
Sexo masculino	-5125,03	333,1629	-15,383	2,88E-49
Imc	339,6139	28,61815	11,86708	6,03E-31
Dependentes	476,8809	137,8941	3,458313	0,000561
Tabagismo (sim)	23842,49	413,4232	57,67091	0
Região sudeste	-1041,41	479,0048	-2,17412	0,029872
Região noroeste	-356,57	476,5868	-0,74818	0,454487
Região sudeste	-957,952	478,2451	-2,00306	0,045374

Onde: Estimate → estimativa do coeficiente; Std. Error → erro padrão; t value → estatística t; Pr(>|t|) → p-valor

O valor de R^2 foi de 0,75 demonstrando uma boa robustez do modelo. Além disso, a estatística F com p-valor < 2,2e-16 confirma que, em conjunto, as variáveis independentes explicam de forma estatisticamente significativa a variação no preço dos planos de saúde. E as variáveis idade, sexo masculino, imc, dependentes e tabagismo foram altamente significativas (***) apresentando PR (>|T|) \<0,001.

O modelo apresentou um coeficiente de determinação (R^2) igual a 0,75, indicando que 75% da variação observada nos preços dos planos de saúde é explicada pelas variáveis independentes incluídas na análise. Esse valor reflete um bom poder explicativo do modelo, demonstrando sua relevância para compreender os fatores que impactam a precificação. A equação estimada pelo modelo será apresentada a seguir.

$$preco = -6957{,}33 + 256{,}90 \cdot idade - 5125{,}03 \cdot 1(sexo\ masculino) + 339{,}61 \cdot IMC + 476{,}88 \cdot dependentes + 23842{,}49 \cdot 1(tabagismo, Sim) - 1041{,}41.1(regiao\ sudeste) - 356{,}57 \cdot 1(regiao\ noroeste) - 957{,}95.1(regiao\ sudoeste)$$

Foi elaborado um boxplot (Figura 1) para representar a distribuição dos preços dos planos de saúde de acordo com a idade, agrupada em janelas de cinco anos. O gráfico mostra, de forma clara, a mediana, a dispersão dos valores e a presença de outliers em cada faixa etária. Observa-se que, conforme a idade aumenta, há uma tendência de elevação no preço médio, além de maior variabilidade dos valores, indicando que a idade exerce influência direta na precificação dos planos, ratificando sua significância estatística ao modelo.

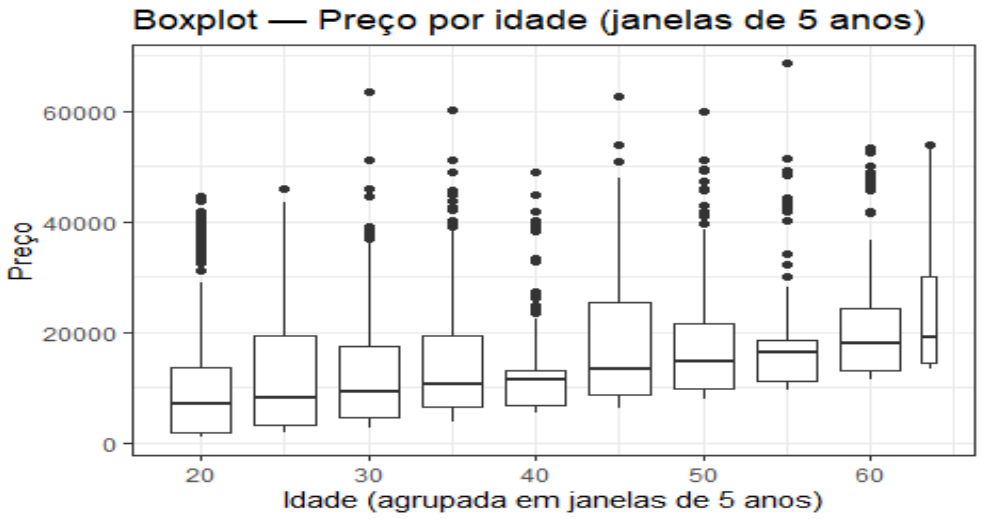

Figura 1: Distribuição do preço dos planos de saúde por faixas etárias (boxplot em janelas de 5 anos).

Uma regressão logística foi usada para modelar e prever a probabilidade de uso do plano (Uso=1) a partir de variáveis como idade, sexo, IMC, nº de dependentes e tabagismo. Os resultados estimados da regressão logística estão apresentados na Tabela 3.

Tabela 3: Estimativas e coeficientes do modelo de regressão logística para a utilização do plano de saúde

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-31,5942	2,202944	-14,3418	1,2E-46
Idade	0,337589	0,023672	14,26084	3,84E-46
Sexo masculino	-0,10149	0,233931	-0,43384	0,664402
Imc	0,49281	0,037024	13,31064	2,01E-40
Dependentes	0,226345	0,102487	2,208515	0,027208
Tabagismo (sim)	1,624756	0,313076	5,189658	2,11E-07

Onde: Estimate → estimativa do coeficiente (log-odds); Std. Error → erro padrão; z value → estatística do teste z; Pr(>|z|) → p-valor associado

Com base nos coeficientes estimados pelo modelo de regressão logística, foi possível montar a equação (n) que descreve a probabilidade de utilização do plano de saúde em função das variáveis independentes analisadas.

$$n = -31.5942 + 0.33760{,}3376 \cdot idade - 0{,}1015 \cdot 1(sexo\ masculino) + 0{,}4928.IMC + 0{,}2263 \cdot dependentes + 1{,}6248.1(tabagismoSim)$$

Além da regressão logística, foi realizada a análise dos Odds Ratios (OR) com intervalo de confiança de 95% (IC95%). Essa análise é utilizada para transformar os coeficientes do modelo logístico em medidas de associação mais intuitivas, indicando quantas vezes aumenta (ou diminui) a chance de utilização do plano de saúde conforme a variação em cada variável explicativa (Tabela 4).

Tabela 4: Razões de chances (Odds Ratios - OR) com intervalos de confiança de 95% para os fatores associados à utilização do plano de saúde.

Variavel	OR (IC95%)	Destaque
Idade	**1.40 (1.34–1.47)**	✓
Sexo (masculino vs feminino)	0.90 (0.57–1.43)	
IMC	**1.64 (1.53–1.77)**	✓
Dependentes	**1.25 (1.03–1.54)**	✓
tabagismo	**5.08 (2.78–9.50)**	✓

Esses achados reforçam que tabagismo, idade e IMC são variáveis fortemente associadas à utilização do plano, com relevância prática e estatística. Quando a OR>1 indica aumento das odds de uso por unidade do preditor. Cada OR, então, quantifica o tamanho do impacto de uma variável. Percebe-se que, nesse caso, o OR com maior impacto ao uso do plano é o tabagismo (OR≈5), ou seja, pessoas que fumam apresentam aproximadamente 5 vezes mais chances de uso do plano de saúde.

Além disso, foi realizada uma classificação entre pessoas acima (e abaixo) de 60 anos. E, posteriormente, uma avaliação expressa risco/probabilidade (prob_uso) e risco em decisão classificatória (uso_prev), sendo comparada ao rótulo observado ‘uso’ para calcular matriz de confusão, acurácia, sensibilidade e especificidade.

A matriz de confusão indica desempenho elevado do classificador considerando a classe positiva = 1 (uso). Os resultando mostraram uma acurácia de 96,64% o que denota concordância “quase perfeita” além do acaso. A matriz de confusão é observada na Tabela 5 e confirma um bom equilíbrio entre acertos nos positivos e nos negativos.

Tabela 5: Matriz de Confusão da Classificação do Uso do Plano de Saúde

	Referência = 0	**Referência = 1**
Predição = 0	853	24
Predição = 1	21	440

A Tabela 5 apresenta a matriz de confusão em contagens, onde as colunas (“Referência”) correspondem aos rótulos observados e as linhas (“Predição”) às classes atribuídas pelo modelo. Observam-se 853 verdadeiros negativos **e** 440 verdadeiros positivos**,** além de 21 falsos positivos **e** 24 falsos negativos**,** totalizando N = 1.338. Esses resultados indicam que o classificador acerta majoritariamente tanto os casos sem uso (0) quanto os com uso (1), com erros equilibrados entre falsos positivos e falsos negativos, sugerindo bom desempenho e ausência de viés sistemático para uma das classes.

Por fim, no modelo avaliado, a sensibilidade (0,948) indica a capacidade de identificar corretamente os casos positivos (uso do plano), enquanto a especificidade (0,976) reflete a habilidade de reconhecer os negativos (não uso). Esses valores altos e equilibrados mostram que o classificador detecta a maior parte dos usuários sem gerar muitos falsos alarmes entre não usuários, evidenciando bom desempenho com o ponto de corte adotado (0,5).