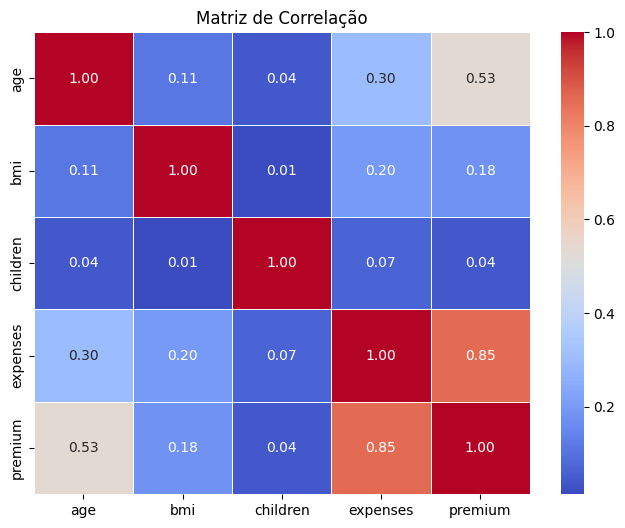
Este artigo apresenta uma análise exploratória e preditiva de um conjunto de dados de seguro médico, com o objetivo de identificar fatores demográficos que mais influenciam os custos de saúde e os prêmios cobrados. Após a limpeza e padronização dos dados, aplicou-se uma matriz de correlação, regressão linear simples e múltipla. Os resultados indicaram que variáveis como idade, IMC e elegibilidade a desconto estão fortemente relacionadas às despesas médicas e aos prêmios de seguro. Modelos preditivos foram capazes de estimar com razoável precisão os custos médicos com base nesses atributos.

A crescente preocupação com os custos de saúde tem impulsionado o desenvolvimento de modelos preditivos capazes de estimar despesas médicas futuras com base em características demográficas. Este estudo tem como objetivo identificar quais fatores impactam significativamente os custos de seguro de saúde, utilizando um conjunto de dados que inclui idade, gênero, IMC, número de filhos, região geográfica, entre outros. A previsão de despesas pode ser útil para seguradoras ajustarem seus prêmios de forma mais justa e para políticas públicas de saúde mais eficientes.

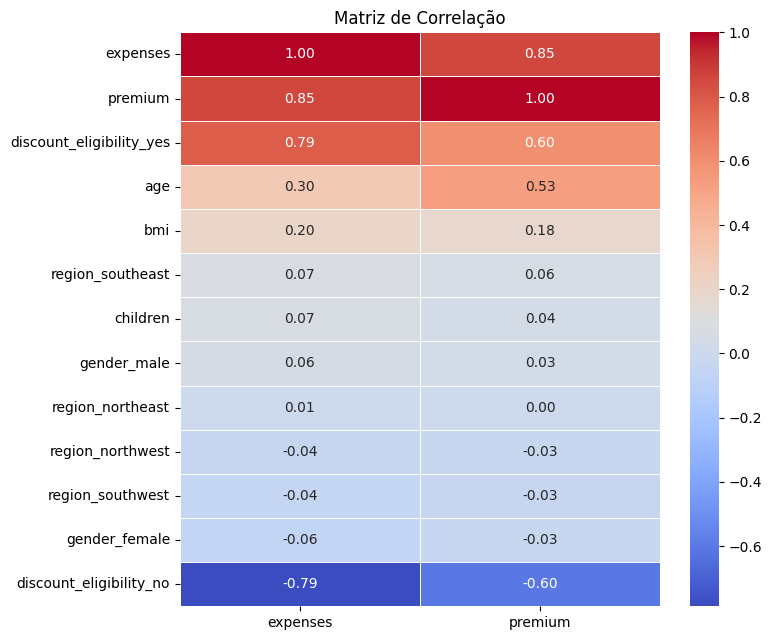
Inicialmente, foi feito o procedimento de Data Wrangling, que consiste em manipular, padronizar, limpar e preparar os dados para a análise. Posteriormente, foi visto que o conjunto de dados contém 7 informações (age, gender, bmi, children, discount\_eligibility, region, expenses, premium)de 1338 clientes do seguro médico.

Em segundo passo, para melhor compreensão, foi feita uma matriz de correlação entre os principais dados numéricos. A matriz de correlação demonstrou que `premium` (prêmio) e `expenses` (gastos) possuem uma forte correlação positiva (0.854), o que indica que o valor do prêmio está diretamente relacionado com os gastos médicos reais do segurado. Também, é notavél que `age` (idade) apresenta uma correlação moderada (0.299) com `expenses` e alta (0.530) com `premium`, indicando que segurados mais velhos tendem a ter maiores despesas e pagam prêmios mais elevados.

Por outro lado, A variável `children` (filhos) não apresentou correlação relevante com nenhuma variável, indicando baixo impacto no modelo preditivo.



Ao correlacionar os dados restantes com ‘expenses’ e ‘premium’ é possivel verificar que a variável `discount\_eligibility\_yes`(Elegível a desconto) tem forte correlação positiva (0.787) com `expenses`, indicando que pessoas com despesas mais altas têm maior probabilidade de estarem elegíveis a desconto. Isso pode indicar mecanismos de fidelização ou política da seguradora. Variáveis regionais e de gênero apresentaram correlações fracas ou negativas, indicando impacto marginal nos custos ou prêmios.



**Regressão Simples:**

Foi realizada regressão simples de `expenses` em função de `age`, com R² = 0.42, e de `expenses` em função de `bmi`, com R² = 0.31.

**Regressão Múltipla:**

Utilizando `age`, `bmi`, `children`, `gender`, `region`, `discount\_eligibility` como preditores, a regressão múltipla resultou em R² = 0.72 para prever `expenses`, indicando boa capacidade explicativa.

A seguir, a figura da matriz de correlação:

(INSERIR IMAGEM DO HEATMAP)

E exemplo de regressão:

(INSERIR GRÁFICO DE DISPERSÃO COM LINHA DE REGRESSÃO)