Disciplina:	Ciência de Dados
Data:	14 de outubro de 2025
Título:	Entrega da primeira etapa do projeto final
Alunos:	Christopher Eduardo Zai, Thiago Ramos Velozo
Repo. Git:	https://github.com/ttrrv/cdpf
Fonte do Dataset:	https://dados.gov.br/dados/conjuntos-dados/caracteristicas-dos-produtos-da-saude-suplementar

Tamanho Estimado do Dataset: Aproximadamente 161 mil registros, totalizando 72,5 MB (original). Devido a limitações de processamento, será utilizada uma amostra de 20 mil registros do dataset original.

ESTRUTURA DO DATASET	
Nome do Campo	**Tipo de Dado**
ID_PLANO	Numérico/Inteiro
CD_PLANO	Texto/Alfanumérico
NM_PLANO	Texto
REGISTRO_OPERADORA	Texto/Numérico
RAZAO_SOCIAL	Texto
GR_MODALIDADE	Texto
PORTE_OPERADORA	Texto
VIGENCIA_PLANO	Texto/Caractere
CONTRATACAO	Texto
GR_CONTRATACAO	Texto
SGMT_ASSISTENCIAL	Texto
GR_SGMT_ASSISTENCIAL	Texto
LG_ODONTOLOGICO	Numérico/Binário
OBSTETRICIA	Texto
COBERTURA	Texto
TIPO_FINANCIAMENTO	Texto
ABRANGENCIA_COBERTURA	Texto
ID_GEO_COBERTURA	Texto/Alfanumérico
FATOR_MODERADOR	Texto
ACOMODACAO_HOSPITALAR	Texto
LIVRE_ESCOLHA	Texto

SITUACAO_PLANO	Texto
DT_SITUACAO	Data (YYYY-MM-DD)
DT_REGISTRO_PLANO	Data (YYYY-MM-DD)
DT_ATUALIZACAO	Data (YYYY-MM-DD)

Contexto e Justificativa do Problema: O mercado de saúde suplementar brasileiro, regulado pela Agência Nacional de Saúde Suplementar (ANS), lida com a constante estabilidade e descontinuidade de produtos. Um plano cancelado ou inativo representa um risco direto para o beneficiário (que precisa migrar para outro plano) e um desafio regulatório para a ANS.

Relevância e Impacto Prático: A capacidade de prever a inatividade (cancelamento) de um plano de saúde é de grande valor estratégico. Para a ANS, um modelo preditivo pode servir como um sistema de alerta precoce, priorizando a fiscalização e a intervenção em operadoras cujos planos apresentem alto risco de descontinuidade. Para as Operadoras, a análise dos fatores de risco pode guiar a criação de produtos mais sustentáveis e a gestão proativa de portfólios.

Beneficiários: A ANS, as Operadoras (em gestão de risco) e, em última instância, os milhões de beneficiários que buscam estabilidade e segurança nos seus contratos de saúde.

Diferenciação: Diferente de estudos que apenas analisam o histórico de reclamações ou dados financeiros, este projeto foca na estrutura contratual e nas características do produto (FATOR_MODERADOR, COBERTURA, ABRANGENCIA_COBERTURA) como preditores primários da estabilidade do plano, utilizando feature engineering em dados temporais para capturar a idade e o ciclo de vida do produto.

Perguntas de Pesquisa

Classificação Predial: É possível predizer a situação de um plano de saúde (Ativo ou Cancelado) com alta acurácia, utilizando as características contratuais, de cobertura e o perfil da operadora (ex: porte e modalidade)?

Importância de Variáveis: Quais fatores (ex: tipo de contratação, abrangência geográfica e tipo de financiamento) têm o maior peso na determinação da situação (estabilidade/instabilidade) de um plano de saúde no Brasil?

Hipóteses Testáveis

H1 (Relação com Variável Temporal - Numérica): A idade do plano (diferença em dias entre a data atual e a DT_REGISTRO_PLANO) é inversamente proporcional à probabilidade de cancelamento: planos mais antigos tendem a ser mais estáveis e ter menor risco de inatividade.

H2 (Relação com Variáveis Categóricas): Planos com abrangência de cobertura menor (Municipal ou Grupo de Municípios) e de contratação Coletivo por Adesão apresentam uma probabilidade significativamente maior de inatividade/cancelamento do que planos Nacionais ou Coletivos Empresariais.

H3 (Modelagem Preditiva): Um modelo de Ensemble Learning (ex: Gradient Boosting ou Random Forest), que consegue capturar interações complexas entre features, obterá uma métrica de desempenho (ex: F1-Score) superior na predição da SITUACAO_PLANO em comparação com um modelo de classificação linear (ex: Regressão Logística).