Predição de Cancelamento e Análise de Ciclo de Vida de Planos de Saúde: Uma Abordagem Baseada em Dados da ANS

Disciplina: Ciência de Dados: Análise de Dados Aplicada (UTFPR) **Data:** 01 de dezembro de 2025 **Autores:** Christopher Eduardo Zai, Thiago Ramos Velozo

1. Resumo Executivo

Este projeto investigou os fatores determinantes para o cancelamento de planos de saúde no Brasil, utilizando dados públicos da Agência Nacional de Saúde Suplementar (ANS). O estudo percorreu o ciclo completo de ciência de dados, desde a integração e limpeza de dados brutos até a modelagem preditiva avançada. Através de uma análise exploratória (EDA) e testes de hipóteses, identificou-se que a "idade do plano" e a "modalidade de contratação" são preditores críticos de estabilidade. Na etapa de modelagem, comparou-se a eficácia de algoritmos lineares versus baseados em árvores. O modelo **Random Forest** superou a Regressão Logística, atingindo um **F1-Score de 0.87**, confirmando a hipótese de que relações não-lineares governam o comportamento de cancelamento. As descobertas sugerem que operadoras devem focar estratégias de retenção em planos recentes e na modalidade "Coletivo por Adesão".

2. Introdução e Definição do Problema

O mercado de saúde suplementar brasileiro enfrenta desafios de sustentabilidade, onde a alta rotatividade (*churn*) de planos impacta a previsibilidade financeira das operadoras e a continuidade do cuidado aos beneficiários. O problema central abordado neste trabalho é a identificação antecipada de planos com alto risco de inatividade.

As perguntas norteadoras da pesquisa foram:

1. **Classificação Preditiva:** É possível predizer a situação de um plano (Ativo ou Cancelado) com alta acurácia utilizando apenas características cadastrais e contratuais?
2. **Fatores de Risco:** Quais variáveis (ex: abrangência geográfica, tipo de contratação) exercem maior peso na decisão de cancelamento?

2.1 Hipóteses Testadas

- **H1 (Temporal):** A idade do plano é inversamente proporcional ao risco; planos mais antigos tendem a ser mais estáveis.

- **H2 (Segmentação):** Planos com abrangência restrita (Municipal) e contratação "Coletivo por Adesão" apresentam taxas de cancelamento significativamente maiores.
- **H3 (Modelagem):** Modelos de *Ensemble* (Random Forest) superam modelos lineares (Regressão Logística) na métrica F1-Score, devido à capacidade de capturar interações complexas entre as variáveis.

3. Metodologia

O pipeline analítico foi desenvolvido em Python e SQL (DuckDB), estruturado em três fases principais: Integração/Limpeza, Análise Exploratória e Modelagem.

3.1 Fonte de Dados e Amostragem

Os dados foram extraídos do conjunto "Características dos Produtos da Saúde Suplementar" (dados.gov.br). Devido ao volume original (~161 mil registros) e limitações computacionais, utilizou-se uma amostragem aleatória estratificada de 10.000 a 20.000 registros para desenvolvimento e validação, garantindo representatividade estatística.

3.2 Pré-processamento e Limpeza (Código Python)

Para garantir a qualidade dos dados, aplicou-se um pipeline rigoroso. O trecho de código abaixo ilustra como as datas foram tratadas e a variável alvo foi normalizada:

```
None
# Pipeline de Limpeza de Dados
# 1. Tratamento Temporal: Conversão para datetime
date_cols = ['DT SITUACAO', 'DT_REGISTRO_PLANO']
for col in date_cols:
    df_clean[col] = pd.to_datetime(df_clean[col],
errors='coerce')

# 2. Normalização da Variável Alvo (Ativo vs Cancelado)
# Removemos status intermediários como 'Suspenso' para focar
na predição binária
df_clean['SITUACAO_PLANO'] = df_clean['SITUACAO_PLANO'].apply(
    lambda x: 'Ativo' if 'ATIVO' in str(x).upper()
    else ('Cancelado' if 'CANCELADO' in str(x).upper() else
'Outro')
)
df_clean = df_clean[df_clean['SITUACAO_PLANO'].isin(['Ativo',
'Cancelado'])]
```

3.3 Feature Engineering (Engenharia de Recursos)

Foi criada a variável `IDADE_PLANO_DIAS` (Data da Situação - Data de Registro), fundamental para testar a H1. Esta transformação permite que o modelo entenda a "maturidade" do contrato.

```
None
# Criação da Feature 'Idade do Plano'
df_clean['IDADE_PLANO_DIAS'] = (
    df_clean['DT SITUACAO'] - df_clean['DT_REGISTRO_PLANO']
).dt.days

# Correção de inconsistências (datas negativas ajustadas para
0)
df_clean['IDADE_PLANO_DIAS'] =
df_clean['IDADE_PLANO_DIAS'].apply(lambda x: max(x, 0))
```

3.4 Análise Exploratória com SQL

Utilizou-se o **DuckDB** para realizar consultas SQL analíticas diretamente sobre o DataFrame Pandas. Isso permitiu validar rapidamente a hipótese H2 (Segmentação).

Consulta SQL: Taxa de Cancelamento por Tipo de Contratação

```
None
SELECT
    CONTRATACAO,
    COUNT(*) as total_planos,
    SUM(CASE WHEN SITUACAO_PLANO = 'Cancelado' THEN 1 ELSE 0
END) as total_cancelados,
    ROUND(SUM(CASE WHEN SITUACAO_PLANO = 'Cancelado' THEN 1.0
ELSE 0 END) / COUNT(*) * 100, 2) as taxa_cancelamento
FROM planos_clean
GROUP BY 1
ORDER BY 4 DESC
```

3.5 Estratégia de Modelagem

- **Codificação:** Variáveis categóricas foram transformadas via *One-Hot Encoding*.
- **Divisão:** Split de treino/teste (80/20).
- **Modelos:**

1. **Baseline:** Regressão Logística (para testar linearidade).
2. **Desafiante:** Random Forest Classifier (para capturar não-linearidades).

```
None
# Configuração dos Modelos
models = {
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "Random Forest": RandomForestClassifier(n_estimators=100,
random_state=42)
}
```

4. Resultados das Análises

4.1 Validação de Hipóteses (EDA)

A análise exploratória confirmou as hipóteses iniciais:

- **Confirmação da H1:** Observou-se uma correlação clara entre a idade do plano e a estabilidade. A média de idade dos planos "Cancelados" mostrou-se significativamente inferior à dos planos "Ativos", indicando que a mortalidade dos produtos ocorre predominantemente nos primeiros anos de vigência (*early churn*).
- **Confirmação da H2:** A segmentação por tipo de contratação revelou que "Coletivo por Adesão" possui as maiores taxas relativas de cancelamento, enquanto planos "Empresariais" demonstraram maior retenção. Geograficamente, planos de abrangência "Municipal" mostraram-se mais voláteis que os "Nacionais".

4.2 Performance dos Modelos (H3)

A hipótese H3 foi validada com robustez. O modelo Random Forest apresentou desempenho superior em todas as métricas relevantes, especialmente no F1-Score, que é a média harmônica entre precisão e recall.

Tabela 1: Comparativo de Métricas no Conjunto de Teste

Modelo	Acurácia	Precisio n	Recal l	F1-Scor e
Regressão Logística (Baseline)	0.76	0.75	0.68	0.71
Random Forest (Final)	**0.89**	**0.88**	**0.86**	**0.87**

O Random Forest demonstrou maior capacidade de generalização, reduzindo significativamente os falsos negativos. Isso indica que a decisão de cancelamento não segue uma fronteira linear simples, mas depende de combinações complexas de fatores.

4.3 Importância das Variáveis (Feature Importance)

A análise de importância das features do Random Forest revelou os principais preditores:

1. **IDADE_PLANO_DIAS:** O fator dominante.
2. **PORTE_OPERADORA:** Operadoras de pequeno porte apresentaram padrões distintos.
3. **ABRANGENCIA_GEOGRAFICA:** A restrição geográfica impacta a percepção de valor e, consequentemente, o cancelamento.

5. Discussão e Limitações

Os resultados apontam que as características intrínsecas do produto (o que é vendido e por quem) são fortes preditores de sua longevidade. O sucesso do Random Forest sugere que existem perfis de risco não-lineares. Por exemplo, a variável `IDADE_PLANO_DIAS` não atua sozinha; seu impacto é modulado pelo `PORTE_OPERADORA`.

Limitações Identificadas e Cenários de Falha:

1. **Ausência de Dados Financeiros:** O modelo baseou-se apenas em dados cadastrais. Variáveis como "Preço do Plano" (Prêmio) ou "Sinistralidade" não estavam disponíveis no dataset público, limitando a capacidade de analisar cancelamentos motivados exclusivamente por preço.
2. **Drift Temporal:** O mercado de saúde muda com a regulação. Modelos treinados com dados históricos muito antigos podem perder precisão frente a novas resoluções normativas da ANS.
3. **Incerteza em Novos Planos:** Para planos com menos de 30 dias de vida, o modelo apresenta maior ruído, pois a variável `IDADE_PLANO_DIAS` ainda não maturou.

6. Conclusão e Recomendações

Com base nos dados, recomenda-se às operadoras e gestores:

1. **Foco no Onboarding:** Como a mortalidade é maior nos planos jovens (H1), estratégias de fidelização devem ser intensificadas nos primeiros 24 meses de contrato.
2. **Atenção ao "Coletivo por Adesão":** Este segmento apresentou alta volatilidade. Revisar as regras de entrada e os reajustes para este grupo pode reduzir o *churn*.
3. **Monitoramento Regional:** Planos de abrangência estritamente municipal exigem acompanhamento próximo da rede credenciada local para evitar a perda de valor percebido pelo beneficiário.

Conclusão Final: O projeto atingiu seus objetivos ao demonstrar que é viável utilizar Machine Learning para antecipar o cancelamento de planos de saúde. A entrega de um modelo com **F1-Score de 0.87** provê uma ferramenta acionável para a gestão estratégica de carteiras na saúde suplementar.

7. Referências Bibliográficas

1. BRASIL. Agência Nacional de Saúde Suplementar (ANS). *Características dos produtos da saúde suplementar*. Disponível em: dados.gov.br. Acesso em: Out. 2025.
2. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
3. Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
4. Scikit-Learn Developers. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.