UNIFECAF

LUIS GUSTAVO DOS SANTOS TALGATTI

MELHORIA DO ATENDIMENTO AO PACIENTE ATRAVÉS DA ANÁLISE DE DADOS NA HEALTHCARE SOLUTIONS

Taboão da Serra

Novembro de 2025

1. Cenário-Problema

A HealthCare Solutions, uma rede de hospitais, está em meio a uma iniciativa de transformação digital visando melhorar a qualidade do atendimento ao paciente. A instituição possui um grande volume de dados de diversas fontes, incluindo registros eletrônicos de saúde (EHRs), dispositivos de monitoramento de pacientes, pesquisas de satisfação e sistemas administrativos.

O desafio central é que esses dados se encontram dispersos e mal organizados, o que impede a extração de insights estratégicos.

A análise de dados é fundamental para reverter esse cenário e impactar positivamente a jornada do paciente. Ao unificar e analisar essas fontes, a HealthCare Solutions pode:

- **Transitar de um cuidado reativo para proativo:** Em vez de apenas tratar os pacientes, podemos prever quais pacientes têm maior risco de readmissão hospitalar.
- **Otimizar a eficiência operacional:** Identificar gargalos em processos, como tempos de espera ou alocação de leitos.
- **Melhorar a experiência do paciente:** Correlacionar dados operacionais (ex: tempo de internação) com os resultados das pesquisas de satisfação, entendendo os principais fatores de atrito.

Este projeto foca especificamente no desenvolvimento de um modelo preditivo para identificar o risco de readmissão em 30 dias, permitindo que a equipe clínica tome ações preventivas.

2. Fontes de Dados Utilizadas

Para simular o desafio, utilizamos quatro conjuntos de dados distintos, representando as fontes de informação da HealthCare Solutions:

1. **Registros Eletrônicos de Saúde (EHRs):** A fonte central de dados. Contém informações demográficas do paciente (idade, gênero), detalhes da admissão (datas, diagnóstico primário) e histórico clínico (comorbidades).
2. **Dispositivos de Monitoramento:** Dados brutos de séries temporais dos sinais vitais dos pacientes durante a internação (ex: frequência cardíaca, saturação de oxigênio).
3. **Pesquisas de Satisfação dos Pacientes:** Dados coletados após a alta, contendo notas (1-5) para quesitos como comunicação da equipe, limpeza e satisfação geral.
4. **Dados Administrativos:** Informações de nível hospitalar, como a região da unidade, número total de leitos e rácio de funcionários.

3. Fundamentos da Ciência de Dados (Etapas do Projeto)

O projeto seguiu rigorosamente o fluxo de trabalho padrão da Ciência de Dados para garantir a qualidade e a interpretabilidade dos resultados:

1. **Coleta de Dados:** Leitura e carregamento dos quatro arquivos CSV em um ambiente de análise (Jupyter Notebook).
2. **Tratamento (Limpeza e Pré-processamento):** Etapa crucial onde identificamos e removemos 270 linhas duplicadas. Convertemos colunas de texto (como datas) para formatos numéricos e tratamos valores ausentes (ex: preenchendo comorbidades nulas com "Nenhuma").
3. **Análise Exploratória (EDA):** Investigação visual dos dados. Geramos gráficos, histogramas e um mapa de calor (heatmap) para identificar padrões. Descobrimos, por exemplo, que "Insuficiência Cardíaca" (Heart Failure) era o diagnóstico com maior taxa de readmissão.
4. **Modelagem Preditiva:** Selecionamos a variável alvo (`readmitted_30_days`) e aplicamos algoritmos de Machine Learning (Regressão Logística e Random Forest) para prever este desfecho.
5. **Visualização:** Utilizamos `matplotlib` e `seaborn` para criar visualizações claras, como as Matrizes de Confusão, que comunicam o desempenho dos modelos.
6. **Interpretação de Resultados:** Analisamos os resultados do modelo. A primeira tentativa (Modelo 1) falhou em prever readmissões devido ao desbalanceamento dos dados. O Modelo 2 (corrigido) foi capaz de identificar 43% dos casos de readmissão. Também analisamos a importância das features, descobrindo quais variáveis mais impactavam a previsão.

4. Aspectos Éticos e Legais (LGPD)

A consideração de aspectos éticos é mandatória em projetos de saúde. Os dados de saúde são classificados como **dados sensíveis** pela Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018).

Em um cenário de produção real, as seguintes medidas seriam essenciais:

- **Anonimização:** Todos os dados pessoais (nome, CPF, patient_id) devem ser completamente anonimizados (ex: usando técnicas de hashing) antes de serem disponibilizados para a equipe de ciência de dados.
- **Princípio da Finalidade:** Os dados só podem ser usados para o fim acordado: melhorar o atendimento ao paciente. É eticamente proibido usar esses dados para criar modelos que discriminem pacientes (ex: por gênero, região) ou para fins comerciais (ex: precificação de planos de saúde).
- **Segurança:** Os dados devem ser armazenados em um ambiente seguro, com acesso restrito e rastreável.

Para este projeto acadêmico, utilizamos dados simulados (mock data) que não possuem identificação com pessoas reais, garantindo o cumprimento desses princípios.

5. Levantamento de Requisitos (Perguntas Simuladas)

Para contextualizar a análise, segue uma simulação de 10 perguntas (e respostas) que seriam feitas à equipe de gestão da HealthCare Solutions antes do início do projeto.

1. Pergunta (Objetivo de Negócio): Qual é o principal indicador (KPI) que a diretoria deseja melhorar com este projeto?

- **Resposta Simulada:** "Nosso principal problema é a alta taxa de readmissão não planejada em 30 dias. Está 15% acima da média nacional, gerando custos elevados e penalidades."

2. Pergunta (Usuário Final): Quem irá consumir o resultado deste modelo no dia a dia?

- **Resposta Simulada:** "A equipe de gestão de alta (enfermeiros-chefe e assistentes sociais). Eles precisam de um 'alerta de risco' no sistema, 24h antes da alta programada do paciente."

3. Pergunta (Definição de Sucesso): O que seria um "sucesso" para este projeto em 6 meses?

- **Resposta Simulada:** "Um modelo que consiga identificar corretamente pelo menos 50% dos pacientes que serão readmitidos (Recall). Preferimos ter alguns 'falsos positivos' (alertar pacientes de baixo risco) do que perder um paciente de alto risco."

4. Pergunta (Fonte de Dados - EHR): O campo `primary_diagnosis` (diagnóstico) usa um padrão (como CID-10) ou é texto livre?

- **Resposta Simulada:** "Infelizmente, é um misto. Os médicos mais novos usam o CID-10, mas os antigos digitam por extenso. Vocês terão que padronizar isso."

5. Pergunta (Fonte de Dados - Dispositivos): Com que frequência os dados dos dispositivos (SpO2, batimentos) são coletados?

- **Resposta Simulada:** "Os monitores novos coletam a cada 5 minutos, mas os antigos só registram quando o enfermeiro mede manualmente e digita no sistema. A granularidade será inconsistente."

6. Pergunta (Fonte de Dados - Satisfação): Qual é a taxa de resposta da pesquisa de satisfação?

- **Resposta Simulada:** "É boa, cerca de 90% dos pacientes respondem, pois fazemos via tablet antes da alta."

7. Pergunta (Validação de Dados): Existem readmissões que são *planejadas* (ex: quimioterapia, cirurgias em etapas)?

- **Resposta Simulada:** "Sim, boa pergunta. Precisamos que o modelo preveja apenas readmissões *não planejadas*. Vocês conseguem filtrar as planejadas usando os `procedure_code` (código de procedimento)?"

8. Pergunta (Viés e Ética): Existem preocupações sobre vieses nos dados atuais?

- **Resposta Simulada:** "Sim. Nossos hospitais em regiões de baixa renda tendem a ter pacientes com comorbidades mais mal documentadas. Precisamos garantir que o modelo não penalize esses pacientes."

9. Pergunta (Integração): Onde este modelo rodará? Qual é a infraestrutura de TI?

- **Resposta Simulada:** "Temos um Data Warehouse na nuvem (Azure). O ideal é que o modelo seja 'deployado' como uma API que o nosso sistema de EHR possa consultar em tempo real."

10. Pergunta (Interpretação): A equipe médica aceitará um modelo "caixa-preta" (Black Box)?

- **Resposta Simulada:** "Não. Os médicos não confiarão em um alerta se não souberem o 'porquê'. O modelo precisa ser interpretável, indicando *quais fatores* (ex: 'Idade Alta', 'Diagnóstico X') levaram àquela pontuação de risco."