Relatório Estratégico de Inteligência de Dados: Adoção de Telemedicina

Integrantes:
Diogo Julio - RM553837
Vinicius Silva - RM553240
Victor Didoff - RM552965
Matheus Zottis - RM94119
Jonata Rafael - RM552939

Repositório:
https://github.com/vinirex/IA---ML---Sprint1e2

Assunto:

Classificação de padrões de adoção de telemedicina usando Regressão Logística

Conjunto de Dados: TMEDTREND_PUBLIC_250827.csv

Intro

Este relatório apresenta o desenvolvimento de um modelo preditivo capaz de identificar, com **80.93% de precisão**, se um determinado grupo demográfico numa região específica terá uma "Alta Adoção" de telemedicina.

A análise revela que **fatores estruturais (Ano e Localização Geográfica)** são determinantes muito mais fortes do que a demografia individual. O estudo sugere que as barreiras para a telemedicina não são primariamente culturais (idade do paciente), mas sim infraestruturais e regulatórias.

Definição do Problema e Hipótese

- **Objetivo de Negócio:** Determinar se características regionais e demográficas podem prever o nível de utilização de serviços remotos de saúde.
- **Abordagem de Machine Learning:** Classificação Binária Supervisionada.
- **Definição do Alvo (Target):**
    - **Alta Adoção (1):** Pct_Telehealth > Mediana Global (Aprox. 17.76%).
    - **Baixa Adoção (0):** Pct_Telehealth < ou = Mediana Global.

Metodologia Técnica e Código

Utilizou-se a biblioteca **Scikit-Learn** em Python. O processo envolveu limpeza de dados, codificação de variáveis categóricas (One-Hot Encoding) e modelagem via Regressão Logística.

Preparação dos Dados

O código abaixo ilustra a criação da variável alvo e o tratamento de nulos.

```
import pandas as pd
import numpy as np

# 1. Carregamento e Limpeza Inicial
df = pd.read_csv("TMEDTREND_PUBLIC_250827.csv")
df_clean = df.dropna(subset=['Pct_Telehealth']).copy()

# 2. Definição do Limiar (Mediana) e Criação da Variável Alvo Binária
median_threshold = df_clean['Pct_Telehealth'].median()
df_clean['High_Telehealth'] = (df_clean['Pct_Telehealth'] > median_threshold).astype(int)

# 3. Seleção de Features Relevantes
features = ['Year', 'quarter', 'Bene_Geo_Desc', 'Bene_Race_Desc',
            'Bene_Sex_Desc', 'Bene_Age_Desc', 'Bene_RUCA_Desc']
X = df_clean[features]
y = df_clean['High_Telehealth']

print(f"Limiar de corte (Mediana): {median_threshold:.4f}")
```

Engenharia de Features e Treinamento

Optou-se pela **Regressão Logística** devido à necessidade de interpretabilidade dos coeficientes (Odds Ratios).

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Pipeline de Pré-processamento
# - Year: Normalização (StandardScaler) para capturar tendências lineares
# - Categóricos: One-Hot Encoding para transformar estados/raça em vetores numéricos
categorical_cols = ['quarter', 'Bene_Geo_Desc', 'Bene_Race_Desc',
                    'Bene_Sex_Desc', 'Bene_Age_Desc', 'Bene_RUCA_Desc']
```

```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), ['Year']),
        ('cat', OneHotEncoder(drop='first'), categorical_cols)
    ])

# Definição do Modelo
model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', LogisticRegression(solver='liblinear', random_state=42))
])

# Divisão Treino (80%) / Teste (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Treinamento
model.fit(X_train, y_train)
```

Resultados da Modelagem

Métricas de Desempenho

O modelo apresentou um desempenho equilibrado, sem viés significativo para nenhuma das classes.

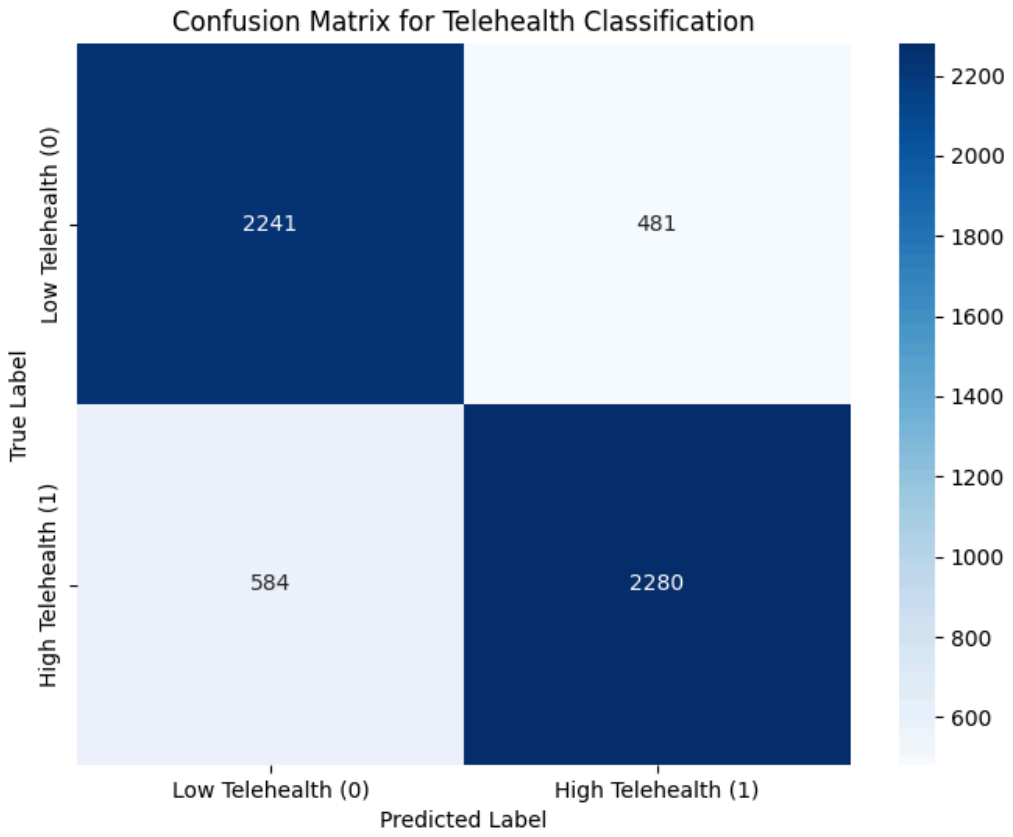
Tabela 1: Resumo das Métricas de Classificação

Métrica	Valor Obtido	Interpretação
Acurácia	**80.93%**	Percentual global de acertos.
Precision (Classe 1)	0.83	Quando o modelo diz que é "Alta", ele acerta 83% das vezes.
Recall (Classe 1)	0.80	O modelo detectou 80% de todos os casos reais de alta adoção.

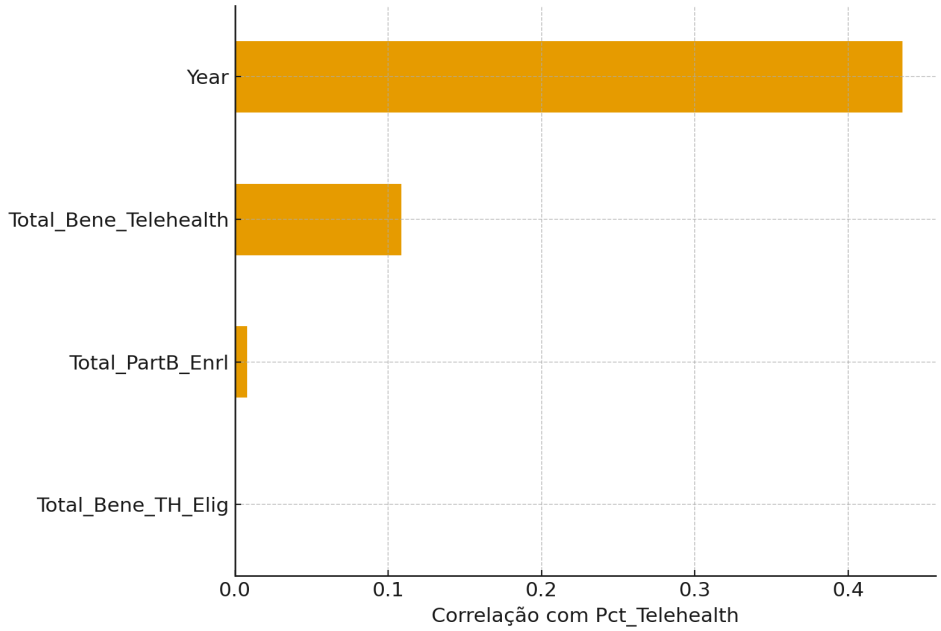
F1-Score	0.81	Média harmônica, indicando robustez.

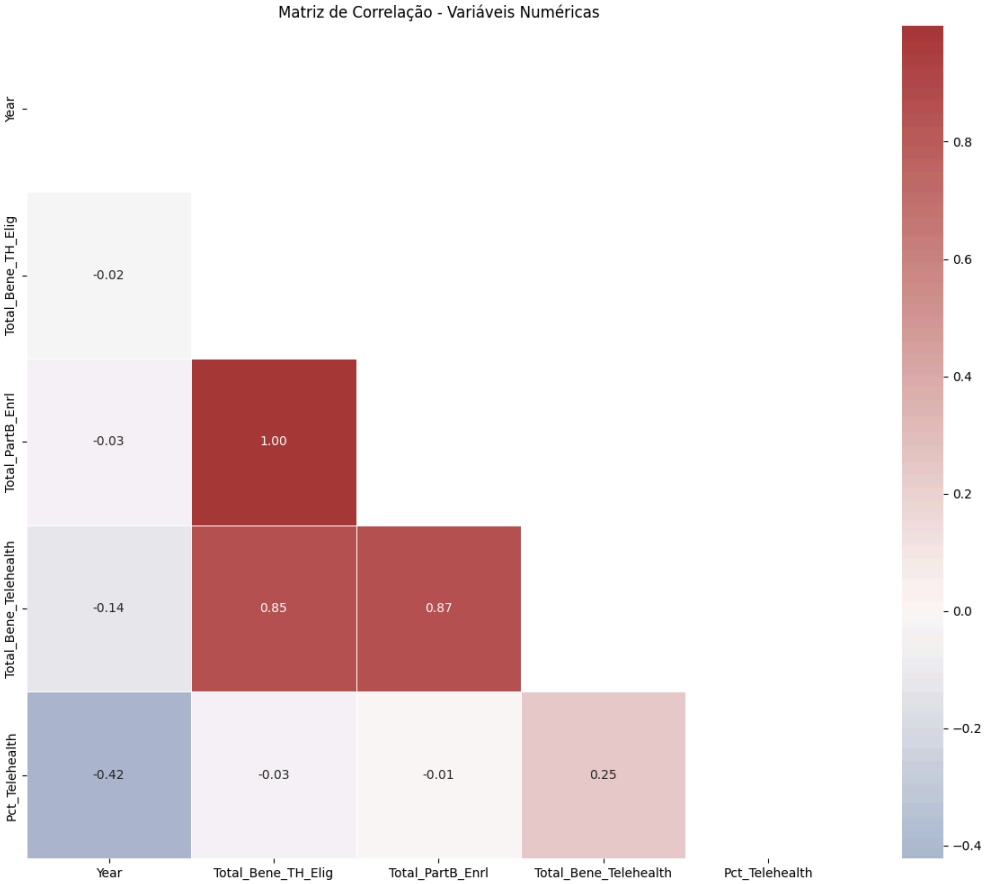
4.2. Matriz de Confusão

A matriz abaixo visualiza onde o modelo acertou e errou. Nota-se um leve aumento nos Falsos Negativos (Tipo II), indicando uma tendência conservadora.



Importância das Features (Coeficientes)





---

5. Análise Estratégica Aprofundada

A análise dos coeficientes gerados pelo código acima revela três pilares críticos de influência:

A. A Força da Tendência Temporal (O Fator "Ano")

A variável `Year` apresentou o coeficiente positivo de maior magnitude.

- **Interpretação:** A adoção da telemedicina não é um evento estático, mas uma mudança de paradigma contínua. O simples passar do tempo (maturidade tecnológica e habituação pós-pandemia) é o maior preditor de sucesso.
- **Insight:** Iniciativas que falharam em 2020 podem ter sucesso agora devido à maior familiaridade do público.

B. A Disparidade Geográfica (O Fator "Estado")

Estados como **Nova Jérsia (NJ)** e **Nova Iorque (NY)** são fortes preditores de alta adoção, enquanto **Minnesota (MN)** e **Dacota do Norte (ND)** predizem baixa adoção.

- **Hipótese Estrutural:** NY e NJ possuem alta densidade urbana e cobertura de banda larga. Já as Dacotas, sendo rurais, enfrentam desafios de infraestrutura. Embora a telemedicina seja ideal para áreas rurais (evita deslocamentos), a falta de internet de alta velocidade atua como um "gargalo oculto".

C. O Paradoxo Demográfico (O Fator "Idade < 65")

O grupo **0-64 anos** apresentou um coeficiente negativo.

- **Contexto Medicare:** Neste dataset, usuários abaixo de 65 anos no Medicare são, tipicamente, Pessoas com Deficiência (PCD) permanente.
- **Análise Crítica:** O fato deste grupo usar *menos* telemedicina que os idosos é alarmante. Sugere que as plataformas atuais podem não ser acessíveis (UX/UI) para pessoas com certas limitações motoras, visuais ou cognitivas, criando uma exclusão digital.

---

6. Conclusão e Recomendações

O modelo prova que a adoção da telemedicina **não é aleatória**. É altamente previsível e depende primariamente de **onde** o paciente está e em que **ano** estamos.

Plano de Ação Recomendado:

1. **Auditoria de Acessibilidade Digital (Foco no Grupo 0-64):**
    - Investigar por que beneficiários com deficiência (Medicare <65) têm menor adoção. Adaptar interfaces para leitores de tela e comandos de voz.

2. **Investimento Direcionado (Infraestrutura vs. Marketing):**
    - Em estados com coeficientes negativos (MN, ND), o problema provavelmente é infraestrutura (sinal), não cultura. O investimento deve focar em banda larga rural, não em campanhas de conscientização.
3. **Benchmarking Regulatório:**
    - Gestores de saúde devem analisar as políticas estaduais de **Nova Iorque**, que se mostrou um *outlier* positivo extremo, para replicar seu modelo de reembolso e regulação em outros estados.