Challenger – Inteligência Artificial / Machine Learning

Tema: Predição de Alta Adoção de Telemedicina (Telehealth)

Curso: Engenharia de Software – FIAP
Turma: 3ESPA

Alunos:

Vitor Pinheiro Nascimento – RM 553693

Gabriel Leão – RM 552642

Miguel Parrado – RM 554007

Matheus Farias – RM 554254

Ano/Semestre: 2025 – 2º semestre

---

1. Introdução

A pandemia de COVID-19 acelerou muito o uso de **telemedicina (telehealth)**, permitindo que pacientes realizassem consultas à distância com médicos e outros profissionais de saúde. Mesmo após o período mais crítico da pandemia, esse modelo de atendimento continua relevante, principalmente em regiões onde o acesso à saúde presencial é mais difícil.

Neste trabalho, o objetivo é utilizar **técnicas de Machine Learning** para prever quais combinações de **características regionais e demográficas** estão associadas a **alta adoção de telemedicina**. Em outras palavras, queremos construir um modelo que receba informações como grupo demográfico, estado, fase e período da pesquisa, e diga se aquele contexto tende a ser de **alta** ou **baixa** utilização de telehealth.

A proposta está alinhada com o desafio apresentado em sala: definir uma variável binária (alta vs baixa adoção) a partir de um indicador de percentual de uso de telemedicina e, a partir disso, treinar um modelo de classificação e analisar suas métricas (acurácia, matriz de confusão, curva ROC, etc.).

---

2. Conjunto de dados e variáveis

Os dados utilizados neste estudo são públicos e foram obtidos a partir de um dataset do **Centers for Disease Control and Prevention (CDC)**, relacionado ao projeto **Household Pulse Survey**, que monitora indicadores de saúde e comportamento da população norte-americana. O arquivo usado contém informações sobre o percentual de adultos que utilizaram telemedicina em um determinado período.

Cada linha do dataset representa uma combinação de:

- **Group**: grupo de análise (por exemplo, por idade, renda, escolaridade etc.);
- **State**: estado ou nível de agregação geográfica (em muitos casos "United States" como total);
- **Subgroup**: subgrupo dentro do grupo principal (por exemplo, faixa etária "18–29 years", "30–39 years", etc.);
- **Phase**: fase da pesquisa (ex.: 3.1, 3.2, ...), que está relacionada ao período de coleta;
- **Time Period**: número da semana/período da pesquisa;
- **Time Period Label**: descrição textual do intervalo de datas (por exemplo, "Apr 14 – Apr 26, 2021");
- **Pct_Telehealth**: percentual estimado de adultos que tiveram consulta com profissional de saúde via telemedicina nas últimas semanas (essa coluna foi renomeada a partir de Value);
- Além disso, o dataset também contém intervalos de confiança e outros campos estatísticos, que não foram utilizados diretamente no modelo.

A variável Pct_Telehealth é o **indicador principal** do nosso estudo, pois a partir dela definimos a classe de **alta** ou **baixa adoção** de telemedicina.

---

3. Preparação dos dados e formulação do problema

A primeira etapa foi realizar o carregamento do arquivo CSV no ambiente do Google Colab, utilizando a biblioteca **pandas**. Em seguida:

1. A coluna Value foi renomeada para **Pct_Telehealth**, para deixar mais claro que se trata do percentual de uso de telemedicina.
2. Linhas com valores ausentes (NaN) nessa coluna foram removidas.

3. Os valores foram convertidos explicitamente para tipo numérico (float), garantindo que o campo pudesse ser usado em cálculos e gráficos.

Em seguida, definimos a variável alvo do problema. O desafio pedido em aula solicita a criação de uma **classe binária** que separa os casos de **alta adoção de telemedicina** dos casos de **baixa adoção**. Para isso:

- Foi calculado o **percentil 75** de Pct_Telehealth, que resultou em um valor de aproximadamente **23,0**.
- Observações com Pct_Telehealth **maior ou igual a 23,0** foram rotuladas como **1 (alta adoção)**;
- As demais observações foram rotuladas como **0 (baixa adoção)**.

Com isso, criamos a coluna **alta_adocao**, que se tornou a variável alvo (**y**) do modelo. Do ponto de vista de Machine Learning, formulamos o problema como uma tarefa de **classificação supervisionada binária**, onde queremos prever alta_adocao a partir das variáveis categóricas do contexto (Group, State, Subgroup, Phase e Time Period).

---

4. Análise exploratória dos dados (EDA)

Antes de treinar o modelo, foi feita uma **análise exploratória de dados (EDA)** para entender melhor a distribuição do uso de telemedicina e o balanceamento das classes.

No histograma da variável **Pct_Telehealth** (Figura 1), observamos que a maior parte dos valores está concentrada entre aproximadamente **12% e 25%**, com poucos casos acima de 30%. Isso indica que, para a maior parte dos grupos e períodos, o uso de telemedicina é moderado, com poucos contextos de adoção extremamente alta.

Também analisamos a distribuição da variável **alta_adocao** (Figura 2). Os resultados mostram que cerca de **75% das observações** ficaram na classe **0 (baixa adoção)** e aproximadamente **25%** na classe **1 (alta adoção)**. Ou seja, há um certo **desbalanceamento** de classes, o que é importante para interpretar as métricas do modelo, principalmente na classe minoritária (alta adoção).

Essas análises iniciais ajudaram a validar a escolha do percentil 75 como limiar para separar alta e baixa adoção, além de mostrar que existe variação suficiente nos dados para treinar um modelo de classificação.

Figura 1 – Histograma da distribuição da variável Pct_Telehealth.

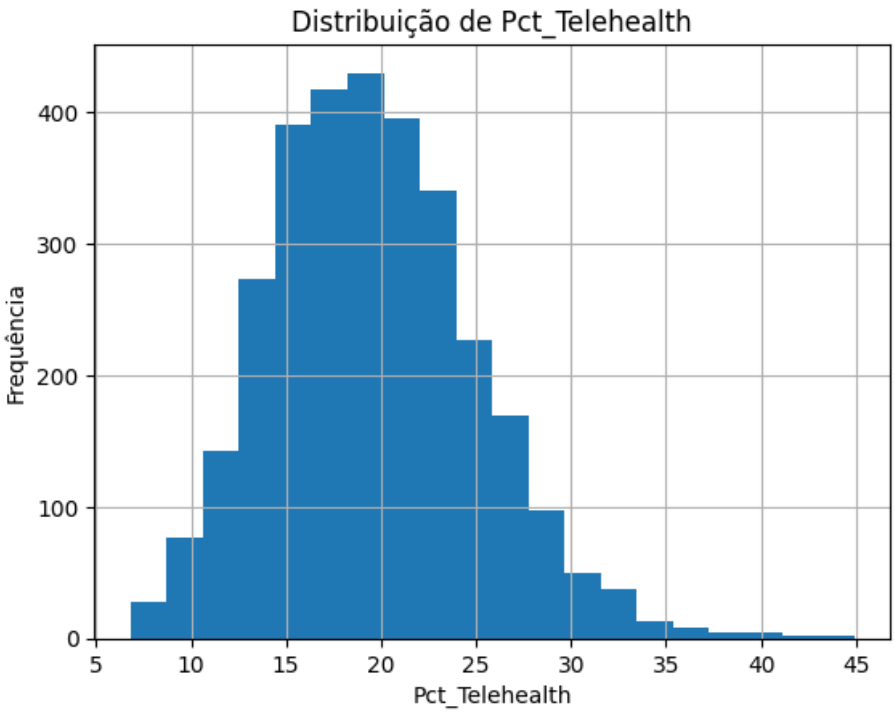
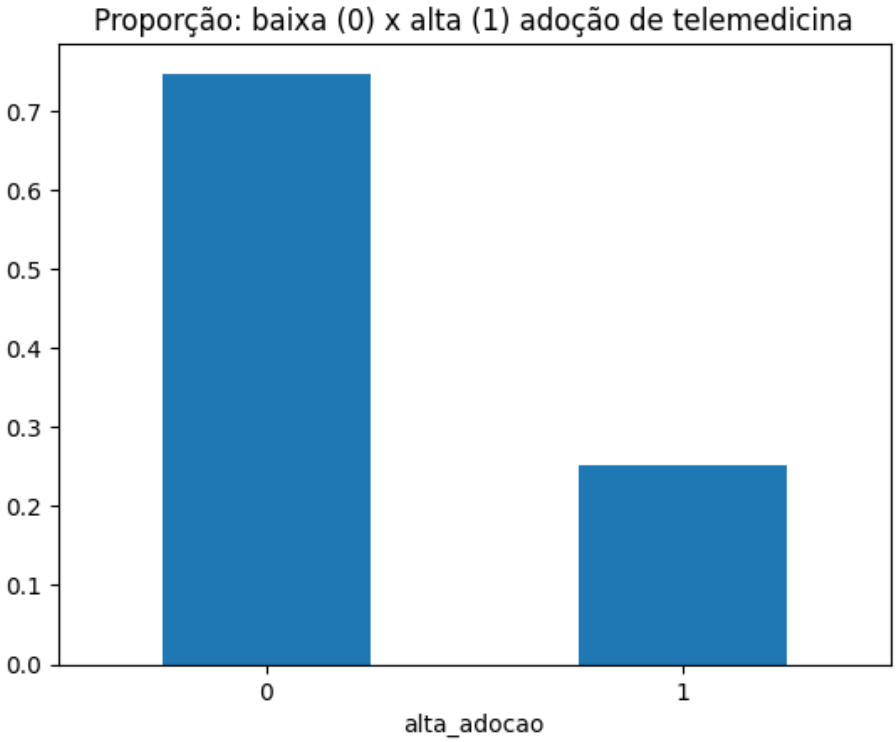
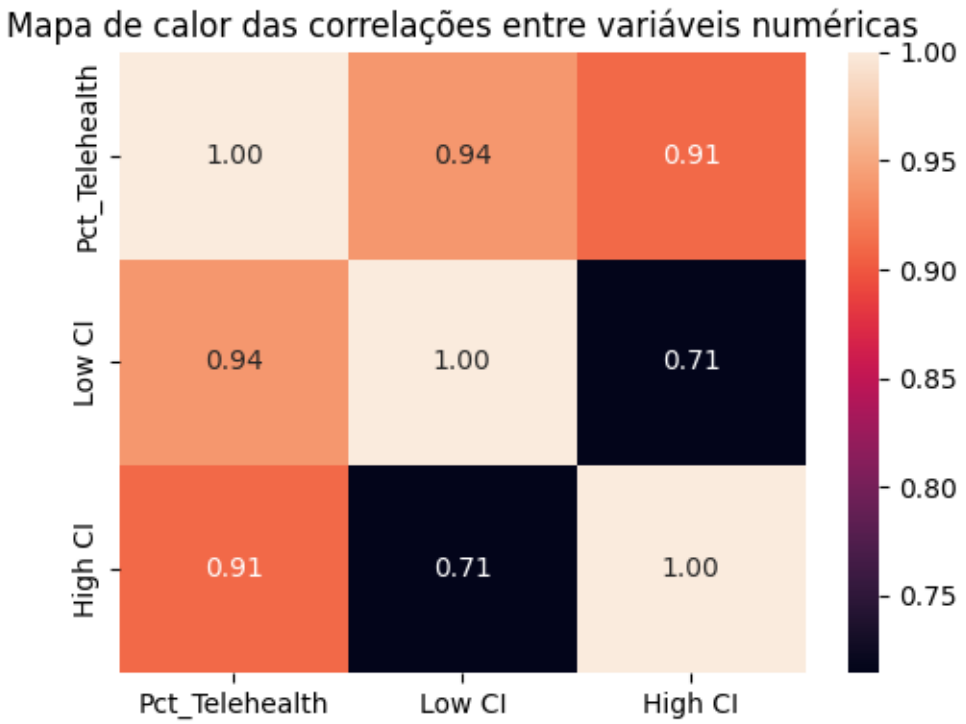


Figura 2 – Distribuição da variável alta_adocao (proporção de baixa vs alta adoção).



Também foi gerado um mapa de calor de correlações entre Pct_Telehealth e os limites inferior e superior do intervalo de confiança. As correlações fortes entre essas variáveis são esperadas, pois todas medem aspectos relacionados à estimativa do percentual de uso de telemedicina.



---

5. Modelo de Machine Learning

Para a modelagem, optamos por utilizar um algoritmo relativamente simples e interpretável, a **Regressão Logística**, implementada com a biblioteca **scikit-learn**.

As variáveis explicativas (**features**) utilizadas foram:

- Group
- State
- Subgroup
- Phase
- Time Period

Como todas essas colunas são categóricas, aplicamos a técnica de **one-hot encoding** (via pd.get_dummies) para transformar cada categoria em um conjunto de variáveis binárias (0 ou 1). Após essa transformação, a matriz de atributos (X) ficou com **3.104 observações** e **158 colunas**. A variável alvo (y) foi a coluna alta_adocao.

O conjunto de dados foi dividido em:

- **70% para treino**,
- **30% para teste**,
  usando a função train_test_split com stratify=y para manter a proporção de classes em ambos os conjuntos.

Em seguida, treinamos o modelo com LogisticRegression(max_iter=1000) e usamos o conjunto de teste para avaliar o desempenho por meio de métricas clássicas de classificação: **acurácia, precision, recall, f1-score**, além da **matriz de confusão** e da **curva ROC com AUC**.

---

6. Resultados

O relatório de classificação obtido no conjunto de teste mostrou os seguintes resultados principais:

- **Classe 0 (baixa adoção de telemedicina)**
  - Precision: **0,85**
  - Recall: **0,93**
  - F1-score: **0,89**
  - Support: 697 observações
- **Classe 1 (alta adoção de telemedicina)**
  - Precision: **0,72**
  - Recall: **0,50**
  - F1-score: **0,59**
  - Support: 235 observações
- **Acurácia global: 0,82** (82%)
- **F1-score ponderado:** aproximadamente **0,81**

Esses valores indicam que o modelo tem um bom desempenho geral, principalmente na classe de baixa adoção, que é a classe majoritária.

A **matriz de confusão** obtida é apresentada na Figura 3. Nela, podemos observar:

- **651 verdadeiros negativos (TN)**: casos de baixa adoção previstos corretamente como baixa;
- **46 falsos positivos (FP)**: casos de baixa adoção previstos como alta;
- **117 verdadeiros positivos (TP)**: casos de alta adoção previstos corretamente como alta;
- **118 falsos negativos (FN)**: casos de alta adoção previstos como baixa.

A **curva ROC** (Figura 4) mostrou um **AUC (Área sob a curva) de aproximadamente 0,861**, o que indica que o modelo possui boa capacidade de distinguir entre os grupos de alta e baixa adoção quando analisamos as probabilidades preditas.

Figura 3 – Matriz de confusão do modelo de classificação de alta vs baixa adoção de telemedicina.

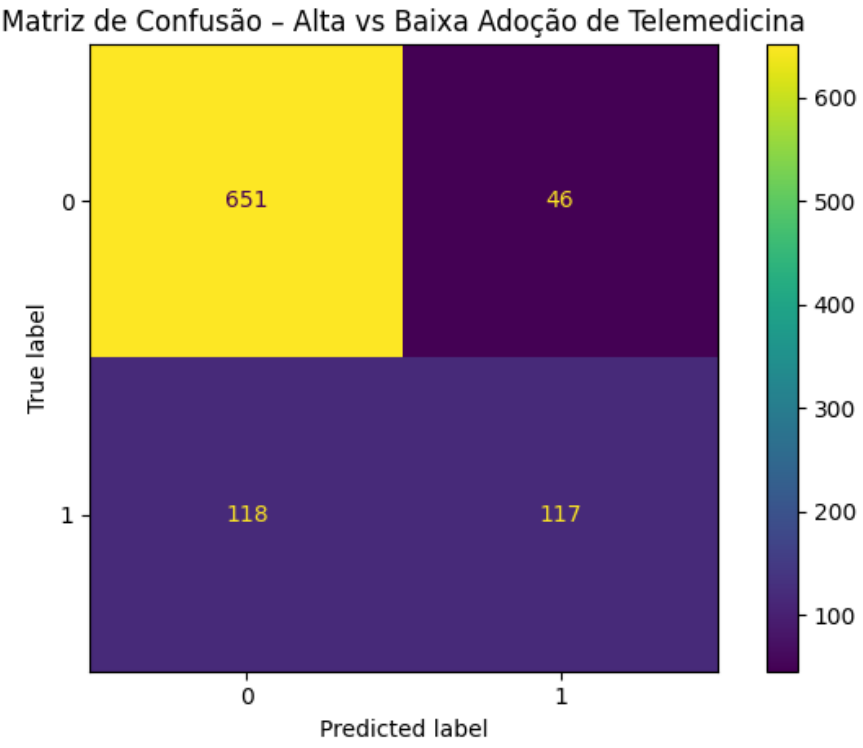
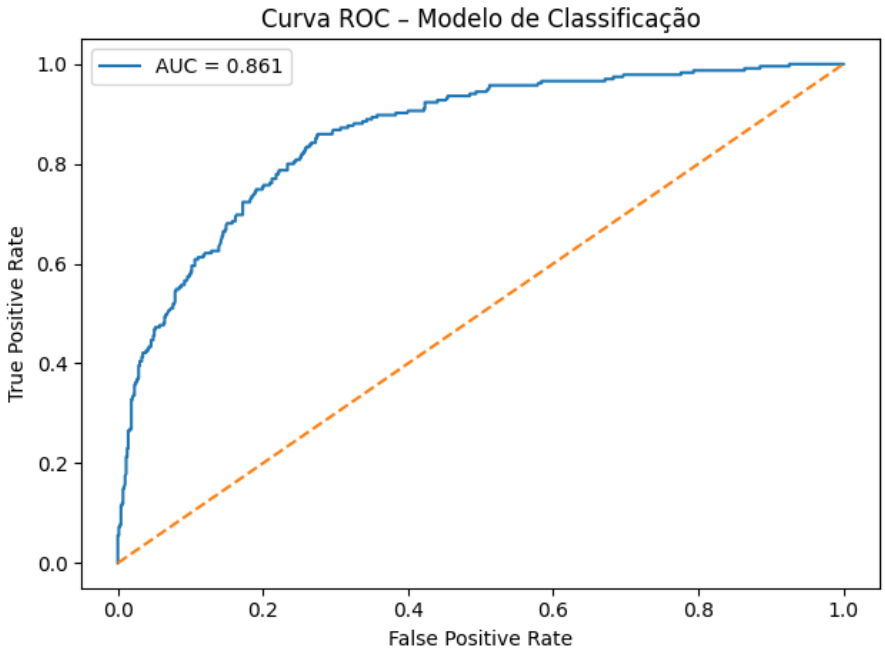


Figura 4 – Curva ROC do modelo de classificação e valor de AUC (0,861).



---

7. Discussão

De forma geral, os resultados indicam que o modelo de Regressão Logística foi capaz de **aprender padrões relevantes** nos dados e atingir uma acurácia global de 82%, com bom equilíbrio entre as classes.

O desempenho na **classe de baixa adoção (0)** foi bastante forte, com f1-score de 0,89, o que mostra que o modelo erra pouco ao classificar contextos em que o uso de telemedicina é mais baixo. Isso é coerente com o fato de essa classe ser majoritária no conjunto de dados (~75% das observações).

Já na **classe de alta adoção (1)**, o modelo apresenta um f1-score de 0,59 e recall de 0,50. Na prática, isso significa que **cerca da metade dos grupos de alta adoção ainda é confundida com baixa adoção**. Esse comportamento também aparece na matriz de confusão, que mostra 118 falsos negativos para 117 verdadeiros positivos.

Essa dificuldade do modelo em capturar todos os casos de alta adoção está diretamente relacionada ao **desbalanceamento de classes** e possivelmente ao fato de que alguns fatores importantes para a adoção de telemedicina **não estão presentes no dataset**, como qualidade da infraestrutura de internet, políticas

locais de saúde digital, campanhas de divulgação ou características específicas dos planos de saúde.

Mesmo assim, o valor de **AUC = 0,861** indica que, se olharmos apenas para as probabilidades geradas pelo modelo, ele consegue separar bem os dois grupos. Isso sugere que, com alguns ajustes (por exemplo, alterar o limiar de decisão, balancear as classes ou testar modelos mais complexos), seria possível melhorar ainda mais o recall da classe de alta adoção sem perder tanto desempenho na classe de baixa adoção.

---

8. Conclusão e trabalhos futuros

Neste Challenge de Inteligência Artificial, utilizamos dados públicos do CDC sobre uso de telemedicina para construir um modelo de classificação capaz de identificar **contextos de alta e baixa adoção de telehealth** a partir de características regionais e demográficas.

O modelo de Regressão Logística alcançou **acurácia de 82%** e **AUC de 0,861**, com excelente desempenho na classe de baixa adoção e desempenho razoável na classe de alta adoção. Isso mostra que, mesmo com um algoritmo relativamente simples, já é possível extrair **insights úteis** que podem apoiar decisões de negócio em áreas como telemedicina, marketing de serviços de saúde e planejamento de políticas públicas.

Como trabalhos futuros, destacamos:

- Experimentar **técnicas de balanceamento de classes** (por exemplo, oversampling da classe 1);
- Testar modelos mais complexos, como **Random Forest** ou **Gradient Boosting**, para comparar desempenho;
- Incluir novas variáveis externas (por exemplo, indicadores de infraestrutura digital por estado, renda média regional, etc.);
- Ajustar o limiar de decisão do modelo para priorizar maior recall na classe de alta adoção, dependendo da estratégia de negócio.

Mesmo sendo um projeto acadêmico, o exercício mostrou na prática como **dados reais + Machine Learning** podem ser usados para entender melhor o comportamento de adoção de telemedicina em diferentes grupos da população.

---

9. Referências

- Centers for Disease Control and Prevention (CDC). *Household Pulse Survey – Telemedicine Use*.
- Documentação oficial do scikit-learn: https://scikit-learn.org
- Material de aula da disciplina de Inteligência Artificial / Machine Learning – FIAP.