CRISP-DM: Projeto Classificador de Bullying

1. BUSINESS UNDERSTANDING

1.1. Determine os objetivos de negócios

1.1.1. Objetivo de negócios

a. Descreva o problema

A fase escolar é uma etapa do desenvolvimento muito importante para o indivíduo. Entender quem se é e a posição que ocupa em um contexto de convivência com muitos outros jovens passando pelo mesmo traz muitos conflitos e, em muitos casos, é possível verificar a violência permeando diversas relações. Com a internet, esse processo se intensificou devido ao anonimato prometido pelas redes.

Em 2018, o Global School-Based Student Health Survey (GSHS) foi realizado na Argentina. O questionário autoaplicável foi entregue aos estudantes e levou em consideração diversos aspectos como peso, situação sócio emocional e familiar, amizades próximas, entre outros. Através desse dataset, a squad Mae C. Jemison tem por objetivo entender o perfil dos jovens e entender os padrões que levam alguém a sofrer ou cometer violência nesses contextos.

b. Impacto do problema

No artigo “Os impactos do bullying na infância e na adolescência”, escrito pela jornalista Maiara Ribeiro para o portal Dráuzio Varella, afirma-se que 38% das escolas brasileiras enfrentam problemas com o Bullying. Para além da violência e do trauma, quem sofre bullying pode desenvolver diversos transtornos e dificuldades na socialização, hábitos destrutivos e da autoimagem como um todo. No entanto, entender o perfil do agressor também é importante, pois ele pode denunciar outros tipos de negligência, como a parental ou emocional, por exemplo. Por isso, entender os fatores que levam a uma situação em que o bullying ocorre e agir de forma preventiva garante a segurança e o bem-estar de uma comunidade como um todo.

c. Questões de negócios

As principais questões de negócios que estamos tentando responder com a análise de dados são:

1. Qual o perfil dos jovens em que o questionário foi aplicado? (Gênero, peso ou se já sofreram algum tipo de violência, por exemplo)
2. Dentre os jovens que afirmam ter sofrido violência, de que formas eles se relacionam? (Se foi violência física ou cyberbullying, com que frequência enfrentam sentimentos de solidão e assiduidade nas aulas, por exemplo).
3. Como é a sua relação com pessoas próximas? (Se eles têm amigos próximos ou a compreensão dos pais, por exemplo).

d. Benefícios em termos de negócios

Ao agrupar e entender a correlação entre fatores socioemocionais, físicos e interpessoais, é possível identificar padrões que podem levar alguém a ser vítima e atuar de forma preventiva em questões de saúde mental e bem-estar.

1.1.2. Background do projeto

a. Unidades de negócios afetadas

Para verificar a acurácia do modelo desenvolvido para o projeto, é necessário envolver a comunidade escolar como um todo: alunos, pais e responsáveis, professores, psicopedagogos e a administração escolar como um todo.

b. Problema em termos gerais

A qualidade e disponibilidade dos dados diz respeito à comunidade escolar na Argentina - no entanto, o contexto social e costumes de uma população também são fatores determinantes para a forma com que os alunos se relacionam entre si.
Além disso, há outros fatores não óbvios que surgem em análises mais detalhadas. Por isso, a interpretação dos dados também irá variar de acordo com quem analisa os dados. Por fim, a ética ao lidar com dados de menores de idade pode dificultar análises mais precisas.

c. Pré-requisitos do projeto

- Motivações: A tomada de decisão baseada em dados e estatísticas para a prevenção de comportamentos danosos trará benefícios para a comunidade acadêmica e o bem-estar dos alunos. Além disso, medidas como o suporte e a validação emocional adequadas para cada fator de risco trará uma personalização dos métodos de prevenção.
- Mineração de dados: Para a realização do estudo, é necessário que haja a coleta, preparação dos dados, aplicação de técnicas de mineração e a avaliação e validação dos modelos e interpretação e ação.

d. Solução usada atualmente

Atualmente, não há um plano de combate ao bullying previamente estabelecido. A administração escolar e equipes psicopedagógicas fazem campanhas, promovem rodas de conversa entre a comunidade sobre saúde mental e buscam incentivar os alunos a se abrirem mais.

e. Vantagens e desvantagens da solução atual

- Vantagens: Buscam engajar mais a comunidade nas pautas de saúde mental e bem-estar. Outra tentativa é de aproximar os pais ou responsáveis do ambiente para buscar uma solução mais adequada.
- Desvantagens: Por conta dos diversos fatores que podem levar alguém a uma situação de vulnerabilidade, uma abordagem mais ampla não é capaz de mitigar as ocorrências de forma efetiva.

f. Objetivos

Classificar indivíduos que sofreram ou não bullying a partir das variáveis presentes.

1.2 Avalie a situação atual

1.2.1. Inventário de recursos

a. Recursos disponíveis para o projeto

i. Pessoal:

Ana Carolina Szczepanski Oliveira, Claisa Lubke, Elis Regina Weiss, Hanna Câmara da Justa, Iris Brandao Pires Linhares, Isabella Stersa de Oliveira, Larissa das Chagas Brum, Lidiane Vicente

ii. Dados: 'Bullying_2018.csv'
iii. Recursos de computação:
Linguagem Python, IDE: Jupyter, Software: Google Colab, Google Meet, Discord, Trello

b. Fontes de dados

Os dados analisados foram disponibilizados no formato CSV e podem ser minerados. Os dados foram anonimizados e não foram utilizados dados de outras fontes. Os dados qualitativos informam sobre:

i. Os dados qualitativos informam sobre:
- A ocorrência (ou não) do bullying
- Em caso de bullying, se foi dentro ou fora da escola
- Sexo
- Se enfrentavam sentimento de solidão
- Se percebem gentileza por parte dos outros estudantes
- Se há compreensão parental sobre a situação
- Se já saíram da escola sem autorização prévia
- Classificação de peso

ii. Os dados quantitativos informam sobre:
- Quantas faltas possuem sem autorização
- Número de amigos próximos
- Quantas vezes já se envolveram em conflitos com violência física
- Idade
- Tome nota dos tipos e formatos de dados.

1.2.2. Requisitos, suposições e restrições

a. Requisitos
i. Conclusão do projeto: 20/12/2024
ii. Dados disponíveis: Dados em arquivo csv, nomeado como Bullying_2018.csv.
iii. Dados sensíveis: Não há dados sensíveis de pessoas
iv. Criar dashboard interativo com indicadores
v. Criar e identificar a previsão de classificação

b. Suposições
i. Expectativa de ver os resultados em
ii. Dados distribuídos como sendo

c. Restrições
i. Dados ausentes: células vazias em várias linhas, abrangendo todas as colunas, exceto a coluna "Record"

1.2.3. Riscos e contingências

a. Riscos

i. Incompletude dos dados: A falta de dados essenciais pode comprometer a precisão da previsão e a qualidade da classificação.
ii. Ambiguidade no produto final: A falta de uma definição clara sobre o produto final pode dificultar a execução de atividades com objetivos bem estabelecidos.

b. Contingências
i. Análise de dados disponíveis: Criar gráficos e realizar análises detalhadas com os dados existentes, buscando extrair o máximo de insights possíveis, mesmo diante da limitação de informações.
ii. Revisão contínua do produto final: Garantir uma definição progressiva e clara do produto final ao longo do processo, alinhando com as partes interessadas para minimizar ambiguidades.

1.3. Objetivos para mineração de dados

1.3.1. Metas de mineração de dados

a. Tipo de problema da mineração de dados
i. Problema de Classificação: O objetivo do modelo é classificar os estudantes em duas categorias: aqueles que probabilisticamente sofrerão bullying na escola (classe "Yes") e os que não sofrerão bullying (classe "No").

b. Modelo a Ser Utilizado:
i. Modelo de Regressão Logística (ou outro modelo de classificação adequado, como Árvore de Decisão ou Random Forest), para prever a probabilidade de um aluno sofrer bullying, com base em variáveis explicativas como características demográficas, comportamentais e sociais.

c. Números reais para os resultados desejados
i. Recall para a Classe "Yes": O modelo deve atingir um recall mínimo de 70% para a classe "Yes", ou seja, o modelo deve ser capaz de identificar corretamente 70% dos alunos que realmente sofrerão bullying. Isso é importante para garantir que o modelo não perca muitos casos de alunos que poderiam ser vítimas de bullying, permitindo que intervenções possam ser tomadas a tempo.
ii. Acuracidade Balanceada (Precision e Recall): Além de focar no recall, deve-se monitorar o equilíbrio entre precisão (precision) e recall, para evitar que o modelo tenha um viés para uma classe em detrimento da outra. O foco será garantir que o modelo tenha uma boa eficiência em identificar vítimas de bullying, sem gerar muitos falsos positivos ou negativos.
iii. Probabilidade de Acerto (Accuracy): O modelo deve alcançar uma probabilidade de acerto pelo menos de 70%, ou seja, a proporção de previsões corretas em relação ao total de previsões realizadas deve ser superior a 70%.

1.3.2. Critérios de sucesso da mineração de dados

a. Métodos para avaliação do modelo

i. *A avaliação do modelo será baseada em uma combinação de métricas quantitativas e métodos subjetivos, garantindo uma análise abrangente do desempenho do modelo.*
    1. *Acuracidade (Accuracy): Medida global da performance do modelo, representando a proporção de previsões corretas (tanto para a classe "Yes" quanto "No") em relação ao total de previsões realizadas.*
       Meta: Acima de 70% de acerto.

    2. *Recall (Sensibilidade) para a Classe "Yes": Mede a capacidade do modelo de identificar corretamente os casos de bullying (classe "Yes"), essencial para garantir que os alunos que realmente enfrentam bullying não sejam negligenciados.*
       Meta: Recall de 70% ou superior para a classe "Yes".

    3. *Precisão (Precision): Avalia a proporção de predições positivas corretas em relação ao total de predições positivas feitas. Isso ajuda a medir a confiabilidade do modelo na previsão da classe "Yes".*
       Meta: A precisão deve ser alta o suficiente para garantir que as intervenções não sejam desnecessárias, mas também deve evitar falsos negativos.

    4. *F1-Score: Uma métrica que combina precisão e recall, oferecendo uma avaliação balanceada entre os dois. Ideal para casos em que é importante minimizar tanto os falsos positivos quanto os falsos negativos.*
       Meta: F1-Score equilibrado que reflete boa performance geral do modelo.

    5. *AUC-ROC (Área sob a Curva ROC): mede a habilidade do modelo em discriminar entre as classes (bullying ou não bullying). A curva ROC ajuda a avaliar a sensibilidade (recall) versus a especificidade do modelo.*
       Meta: AUC superior a 0.80, indicando uma boa discriminação entre as classes.

b. Medidas subjetivas
    i. Avaliar o quão facilmente o modelo pode ser interpretado e compreendido pelos stakeholders, como professores, psicólogos escolares e administradores. A clareza das decisões do modelo e a capacidade de explicar as previsões (como, por exemplo, quais variáveis influenciam a previsão) será crucial para a implementação prática.
    ii. Utilização Prática: Verificar se o modelo é aplicável e útil em situações reais. Isso inclui avaliar a facilidade de integração com os sistemas escolares e a aceitação dos resultados por parte dos profissionais que irão usar as previsões (professores, orientadores, etc.).

iii. Impacto Social e Educacional: Avaliar o impacto do modelo na prevenção de bullying nas escolas. Isso pode incluir feedback qualitativo dos alunos, pais e educadores sobre a percepção de segurança na escola, além da eficácia das intervenções realizadas com base nas previsões do modelo.

iv. Capacidade de Escalabilidade: Avaliar se o modelo pode ser escalado para diferentes escolas ou sistemas educacionais, levando em consideração a variabilidade de dados entre diferentes regiões ou tipos de escolas, mantendo um bom desempenho.

v. Feedback e Adoção: Coletar feedback contínuo das partes interessadas (professores, alunos, psicólogos) sobre a eficácia do modelo e sua aceitação. O modelo será considerado bem-sucedido se for amplamente adotado e se resultar em melhorias tangíveis na segurança escolar.

1.4. Produzir Plano de Projeto

1.4.1. Plano do Projeto

Fase	Data	Responsáveis
Business Understanding	03/12/2024	Ana
Data Understanding	03/12/2024	Iris, Camila
Data Preparation	05/12/2024	Elis
Análise dos Dados	08/12/2024	Claísa, Hanna, Lidiane
4.1. Calcular a matriz de correlação para todas as variáveis numéricas e identificar as correlações mais fortes e mais fracas		Claísa
4.2. Realizar análise exploratória dos dados categóricos, calcular a moda e as variáveis, utilizar recursos visuais e o teste qui-quadrado para avaliar a análise de correspondência		Lidiane
4.3. Plotar histogramas e boxplots para cada tipo de bullying, separados por gênero e idade		Hanna
Modeling	11/12/2024	Hanna, Isa
5.1. Construir um modelo de regressão logística para prever a probabilidade de um indivíduo sofrer bullying na escola, considerando como target: `Bullied_on_school_property_in_past_12_months`		Elis
5.2. Análise das variáveis mais relevantes no modelo		Hanna
5.3. Avaliar a performance do modelo utilizando métricas como Acurácia, Precisão, Recall, e plotar a matriz de confusão		Isa
Evaluation	12/12/2024	Isa
Subir no GitHub - Projeto Final	20/12/2024	Elis Weiss
Criar roteiro do pitch para apresentações	15/12/2024	Larissa
Criar Slide para gravação (20 minutos)	18/12/2024	Larissa
Gravação do Projeto Final	19/12/2024	Hanna, Lidiane, Larissa, Claísa, Elis, Isabella, Ana, Iris
Entrega do Projeto Final	20/12/2024	Elis

1.4.2. Avaliação Inicial de Ferramentas e Técnicas

a. Para o diagnóstico do risco de bullying nas escolas, a técnica de **classificação** é a mais adequada, pois permite categorizar os alunos nas classes "sim" (sofreu bullying) e "não" (não sofreu bullying). Para este projeto, será utilizado o **Logistic Regression Classifier**, que é uma técnica amplamente utilizada para problemas de classificação binária.

b. O **Logistic Regression** é uma abordagem simples e eficaz para prever a probabilidade de ocorrência de um evento (no caso, o risco de sofrer bullying). Ele modela a relação entre as variáveis independentes e a probabilidade de uma resposta binária (1 ou 0) usando a função logística. Além disso, a regressão logística tem a vantagem de ser interpretável, o que facilita a análise dos coeficientes para entender a influência de cada variável.

c. Vantagens da Regressão Logística:
    i. Interpretação Simples**: Facilita a compreensão de como as variáveis explicativas influenciam a probabilidade de sofrer bullying.
    ii. Modelo Probabilístico**: Ao invés de simplesmente classificar, a regressão logística fornece a probabilidade de um aluno sofrer bullying.
    iii. Desempenho em Modelos Lineares**: Funciona bem quando as relações entre as variáveis independentes e a variável dependente são aproximadamente lineares.

d. O modelo será avaliado com base em métricas como **Acurácia**, **Precisão**, **Recall** e **F1-Score**, que ajudarão a verificar a eficácia do modelo em classificar corretamente os alunos que podem sofrer bullying.

2. COMPREENSÃO DOS DADOS

2.1 Relatório de dados inicial

a. Armazenamento dos Dados dos dados estão armazenados em uma pasta na nuvem, utilizando o Google Drive.
b. Importação dos Dados: importação está utilizando a função **`read_csv`** do módulo **`pandas`** para carregar dados de um arquivo CSV em um **DataFrame**
c. Criação do DataFrame: Um DataFrame foi criado a partir de um arquivo do tipo CSV, método pd.read_csv('')
d. Nome do Arquivo: O arquivo contendo os dados possui o nome Bullying_2018.csv.
e. Nome do dataframe original: df_original

2.2 Descrever dados

2.2.2. Análises Iniciais

a. Cópia dataframe:
    i. Criação de Cópia do DataFrame: uma cópia do DataFrame original foi criada para realizar análises iniciais e desenvolver o processo de Data Understanding, garantindo a integridade dos dados originais.
    ii. Nome da cópia do dataframe: df

b. Exibir das Informações do DataFrame: Utilizando o comando df.info(), as seguintes informações foram extraídas:
    i. Número de Registros: O DataFrame contém 56.981 registros.
    ii. Número de Colunas: O DataFrame possui 18 colunas.
    iii. Tipos de Dados:
        1. 1 variável do tipo numérico (int64): A coluna record contém valores inteiros.
        2. 17 variáveis do tipo objeto (string): As demais colunas são de tipo texto, indicando que as variáveis possuem dados categóricos.
    iv. Ausência de Valores Nulos: Não há valores nulos nas colunas, garantindo que os dados estão completos para análise.

c. Verificar a dimensões do DataFrame: Quantidade de Registros e Colunas:
    i. Formato do Conjunto de Dados: O conjunto de dados foi fornecido como um arquivo CSV, contendo 18 colunas e 56.981 registros.

ii. Estrutura do DataFrame: O DataFrame gerado a partir do arquivo CSV possui 18 colunas, sendo 17 variáveis explicativas e 1 variável target.

d. Exibir as colunas, chamar a (1) Função para listar enumerar colunas de um DataFrame verticalmente:
   i. A renomeação das colunas é recomendada para garantir uma nomenclatura mais concisa e padronizada, facilitando a interpretação e análise dos dados. As colunas do DataFrame foram listadas e verificadas, com sugestões de renomeação para maior clareza. A tabela abaixo apresenta os nomes originais das colunas, suas versões renomeadas e uma breve descrição de cada uma:

Nome Original	Nome Renomeado	Descrição
record	record_id	código identificador
Bullied_on_school_property_in_past_12_months	bullied_at_school	Sofreu bullying na propriedade da escola nos últimos 12 meses
Bullied_not_on_school_property_in_past_12_months	bullied_off_school	sofreu bullying fora da propriedade da escola nos últimos 12_meses
Cyber_bullied_in_past_12_months	cyberbullied	Cyberbullying nos últimos 12 meses
Custom_Age	age	Idade
Sex	sex	sexo
Physically_attacked	physically_attacked	Ataque físico
Physical_fighting	physical_fighting	Briga Física
Felt_lonely	felt_lonely	Sentiu-se sozinho
Close_friends	close_friends	Amigos próximos
Miss_school_no_permission	missed_school_no_permission	Faltam à escola sem permissão
Other_students_kind_and_helpful	students_kind_and_helpful	Outros alunos gentis e prestativos
Parents_understand_problems	parents_understand_problems	Os pais entendem os problemas
Most_of_the_time_or_always_felt_lonely	often_felt_lonely	Na maioria das vezes ou sempre se sentiu sozinho
Missed_classes_or_school_without_permission	missed_classes_no_permission	Perdeu aulas ou escola sem permissão
Were_underweight	underweight	Estava abaixo do peso
Were_overweight	overweight	Estava acima do peso
Were_obese	obese	Estava obeso

2.2.2.1. Conclusões e Avaliação dos Dados

Adequação dos Dados aos Requisitos: O conjunto de dados contém variáveis relevantes para a análise de bullying, saúde e fatores sociais. Inicialmente, os registros parecem completos, atendendo aos requisitos estabelecidos. No entanto, será necessário aplicar análises mais detalhadas para avaliar a qualidade dos dados e garantir que atendam a todas as expectativas para as etapas subsequentes da análise.

2.3. Verifique a qualidade dos dados

2.3.1. Dados Duplicados

Foi realizada a verificação de linhas duplicadas no DataFrame, com a análise de registros repetidos. O processo de verificação revelou que não há registros duplicados no conjunto de dados, confirmando a presença de um identificador único para cada linha (campo record)

2.3.2. Dados ausentes

Ao observar as tabelas de frequência de todas as variáveis, foi possível identificar que algumas colunas apresentam uma quantidade significativa de células vazias. Esses

valores ausentes podem ser de diferentes tipos, como NaN, espaços em branco ou até mesmo strings como "?" ou "N/A".

Realizar um tratamento adequado para identificar e substituir os valores faltantes, espaços em branco, NaN, e valores como "?" ou "N/A" por NaN. Isso garantirá uma análise mais precisa e limpa, além de facilitar a aplicação de modelos de machine learning ou outras análises posteriores.

Após a análise dos valores faltantes no conjunto de dados, verificou-se a seguinte distribuição de valores ausentes:

- A variável 'Bullied_on_school_property_in_past_12_months' apresenta 2.17% de valores ausentes. Considerando que esta é a variável target (dependente), recomenda-se remover as células vazias para evitar qualquer viés no modelo de análise.

- As variáveis explicativas, como 'Miss_school_no_permission', 'Other_students_kind_and_helpful', entre outras, possuem valores faltantes, com destaque para as variáveis relacionadas ao peso ('Were_underweight', 'Were_overweight', 'Were_obese') que possuem uma alta porcentagem de dados ausentes (36.73%). Para essas variáveis explicativas, sugere-se a substituição dos valores faltantes por 'No_answer', a fim de manter a consistência e a integridade dos dados para análise.

2.3.3. Outliers

Como o conjunto de dados contém principalmente variáveis categóricas, a análise de outliers não se aplica. Outliers são típicos em variáveis numéricas, onde se busca identificar valores que estão significativamente distantes da maioria dos dados. Sendo assim, neste caso, não realizamos a análise de outliers.

2.3.4. Relatório parcial de qualidades dos dados
Resultados:

Número total de registros: 56.981
Número total de colunas: 18
Número de colunas com valores ausentes: 17
Linhas duplicadas: 0

Esses resultados indicam que, até o momento, os dados não apresentam problemas de duplicação e a quantidade de registros e colunas está dentro do esperado. A análise de valores ausentes será aprofundada nas etapas seguintes para garantir a consistência dos dados.

2.4 Exploração inicial de dados

2.4.1. EDA

2.4.1.1. Análise gráfica

a. Distribuição de frequência da variável target sem substituir NaN:
   - Alta prevalência de não bullying: A grande maioria dos alunos (77%) relatou não ter sofrido bullying na propriedade escolar, o que pode indicar um cenário positivo em relação ao ambiente escolar.

- Proporção significativa de vítimas: Cerca de 21% dos alunos relataram ter sido vítimas de bullying, o que é um número significativo e sugere a necessidade de ações para prevenir e combater o bullying nas escolas.

- Resposta ausente (2%): A taxa de respostas ausentes é baixa, mas ainda assim é importante considerar que alguns alunos podem não ter se sentido confortáveis para responder, o que pode refletir um viés na amostra. Isso também pode ser explorado em análises subsequentes.

- O conjunto de dados apresenta um desbalanceamento considerável entre as classes "No" e "Yes". Com 77% dos registros indicando "No" e 21% indicando "Yes", isso pode causar problemas em modelos de aprendizado de máquina (como modelos de classificação), uma vez que eles podem tender a prever a classe majoritária (No) em detrimento da classe minoritária (Yes).

- Estratégias como sobreamostragem (oversampling) ou subamostragem (undersampling) das classes, ou ainda o uso de técnicas como pesos de classe em modelos de aprendizado de máquina, podem ser aplicadas para lidar com esse desbalanceamento.

b. Visualizar as distribuições de contagem utilizando Histograma para variáveis numéricas e gráfico de Barras para variáveis categóricas:
    i. Frequência de bullying: A maioria dos alunos não relata ter sofrido bullying (seja na escola, fora da escola ou online), mas uma quantidade significativa ainda relatou algum tipo de bullying. Isso indica que o bullying é um problema presente, mas não universal entre os alunos.
    ii. Idade e sexo: A maioria dos alunos está entre 13 e 16 anos, com uma ligeira predominância de meninas (feminino) na amostra.
    iii. Comportamentos de violência: A maioria dos alunos não foi envolvida em ataques físicos ou brigas físicas. Isso sugere que, embora a violência física seja um problema, a maioria dos alunos não está diretamente envolvida.
    iv. Solidão e relações sociais: A maioria dos alunos relatou nunca se sentir solitária, mas uma porção significativa relatou se sentir solitária em algum momento.
    v. Saúde e peso: A maioria dos alunos não estava abaixo do peso ou obesa, mas uma quantidade expressiva foi classificada como acima do peso.

c. Visualizar as distribuições de contagem utilizando Histograma para variáveis numéricas e gráfico de Barras para variáveis categóricas filtrando apenas linhas que sofreram bullying:
    i. **Bullying não ocorre frequentemente em outras situações:** A maioria dos alunos que sofreram bullying na escola também não indicam ter sido vítimas de bullying fora da escola. No entanto, uma proporção considerável ainda reporta ter sofrido bullying em outros contextos, como o ciberbullying (aproximadamente 45% dos alunos).
    ii. **Idade e Sexo:**
        1. **Idade:** A maioria dos alunos que sofreram bullying está entre 13 e 15 anos, com a idade de 14 anos sendo a mais comum. Isso sugere que

adolescentes em faixas etárias mais jovens são mais vulneráveis ao bullying, o que pode ser importante para direcionar intervenções específicas.

2. **Sexo:** A maior parte dos alunos que sofreram bullying são femininas. Esse dado pode refletir diferenças nos padrões de bullying entre sexos ou até mesmo um viés na forma como o bullying é percebido e relatado pelos alunos de diferentes gêneros.

iii. **Ataques Físicos e Brigas:** Um número significativo de alunos relatou ter sido fisicamente atacado uma vez (cerca de 14%) e muitos também indicaram ter participado de brigas físicas (cerca de 16%). As distribuições de "Ataques Físicos" e "Brigas Físicas" mostram que a maioria dos alunos relatou nenhum ataque ou briga, mas uma parte considerável (aproximadamente 10-20%) sofreu ou se envolveu em mais de uma situação.

iv. **Sentimentos de Solidão:** O sentimento de solidão também está presente em uma proporção considerável de estudantes que sofreram bullying, com destaque para os que indicaram sentir-se "algumas vezes" solitários (aproximadamente 30%). As classes "Sempre" e "Na maior parte do tempo" somam cerca de 15%, sugerindo que esses alunos podem estar enfrentando desafios emocionais significativos.

v. **Amigos Próximos e Suporte Social:** A maioria dos estudantes relatou ter 3 ou mais amigos próximos (aproximadamente 60%). No entanto, há uma proporção significativa (cerca de 25%) que tem apenas 1 ou 2 amigos próximos, o que pode indicar uma rede de apoio limitada para esses alunos.

vi. **Falta de Permissão para Faltar à Escola:** A maioria dos alunos que sofreram bullying não faltou à escola sem permissão. No entanto, uma proporção significativa faltou de 1 a 2 dias (cerca de 18%). Esse dado sugere que a experiência de bullying pode afetar a presença escolar, embora a maioria continue frequentando a escola normalmente.

vii. **Ajuda de Outros Estudantes:** A maior parte dos estudantes relatou que outros alunos eram "às vezes" gentis e prestativos, com cerca de 25% também indicando que eram gentis "raramente" ou "na maior parte do tempo". Esse dado pode indicar que a cultura de apoio entre os colegas ainda não é plenamente desenvolvida nas escolas, o que pode ser um fator de risco para o bullying.

viii. **Problemas com os Pais:** A maioria dos alunos indicou que seus pais "nunca" entendem completamente seus problemas (aproximadamente 25%). Esse dado sugere uma lacuna significativa no apoio familiar, o que pode estar relacionado ao comportamento de bullying ou até mesmo a uma falta de recursos ou estratégias para lidar com problemas de bullying.

ix. **Questões de Peso e Saúde:** A grande maioria dos alunos que sofreram bullying não eram obesos ou com sobrepeso (aproximadamente 70-80%). No entanto, há uma proporção significativa de alunos que indicaram não ter respondido à pergunta sobre seu peso (aproximadamente 30-40%), o que pode ser um reflexo de insegurança ou desconforto com o tema.

d. Gráfico de pizza para distribuição de gênero:

i. A distribuição de gênero entre feminino (51,5%) e masculino (47,5%) está muito próxima, indicando um equilíbrio nas representações de gênero na amostra. A diferença é de apenas 4%, o que sugere que os dados são razoavelmente representativos de ambos os gêneros.

ii. Apenas 0,941% das respostas estão marcadas como "No_answer", o que é um valor bastante baixo. Isso indica que a grande maioria dos participantes forneceu uma resposta para a variável de gênero, sugerindo que a pergunta foi clara e que a amostra tem uma boa cobertura de dados válidos.

e. Visualizar grãfico de barras relacinando a coluna Bullied_on_school_property_in_past_12_months' e coluna 'sex':
   i. Prevalência de Respostas sobre Bullying:
      1. 78,9% (43.839 indivíduos) não sofreram bullying.
      2. 21,4% (11.903 indivíduos) relataram bullying, com destaque para mulheres (56,8% dos casos "Yes").
      3. 2,2% (1.239 indivíduos) não responderam.
   ii. Diferenças por Gênero:
      1. Mulheres têm maior probabilidade de relatar bullying (23,5% contra 18,5% dos homens).
      2. Homens apresentam maior proporção de respostas "No" (81,5% contra 76,5% das mulheres).
      3. Grupo "No_answer": Menor proporção de respostas "Yes" (10,2%), sugerindo possível sub-representação ou desconforto em relatar experiências.

2.4.1.2. Correlações

a. Correlação entre Tipos de Bullying:
   i. Bullied_on_school_property_in_past_12_months e Bullied_not_on_school_property_in_past_12_months (0.33).
   ii. Bullied_on_school_property_in_past_12_months e Cyber_bullied_in_past_12_months (0.28).
   iii. Bullied_not_on_school_property_in_past_12_months e Cyber_bullied_in_past_12_months (0.39).

Isso sugere que estudantes vítimas de um tipo de bullying têm maior probabilidade de vivenciar outros tipos.

b. Bullying e Sentimentos de Solidão:
   i. Bullied_on_school_property_in_past_12_months e Felt_lonely (0.17).
   ii. Bullied_not_on_school_property_in_past_12_months e Felt_lonely (0.18).
   iii. Cyber_bullied_in_past_12_months e Felt_lonely (0.20).

Esses dados destacam a conexão entre bullying e sentimentos de isolamento social.

c. Características Relacionadas a Gênero:
   i. Cyber_bullied_in_past_12_months apresenta correlação com Sex (0.11), sugerindo que gênero pode influenciar a exposição ao cyberbullying.

d. Comportamentos Agressivos:
   i. Bullied_on_school_property_in_past_12_months e Physically_attacked (0.18).
   ii. Bullied_not_on_school_property_in_past_12_months e Physically_attacked (0.21).

Isso reflete uma conexão entre bullying e comportamentos físicos agressivos, destacando a necessidade de intervenções preventivas.

e. Suporte Social e Escolar:
    i. Other_students_kind_and_helpful correlaciona-se com:
        1. Miss_school_no_permission (0.29): percepção de menor suporte social pode estar associada a faltas escolares.
        2. Parents_understand_problems (0.31): A percepção de suporte parental é importante na mitigação dos efeitos do bullying.

f. Parâmetros Físicos (peso):
    i. Correlações moderadas entre Were_underweight, Were_overweight e Were_obese (entre 0.71 e 0.77) sugerem uma relação entre os índices de peso corporal, mas sua associação com o bullying apresenta valores baixos, indicando impacto indireto ou menos significativo.

2.4.2. Análise básicas dos dados brutos

a. Agrupando os dados para somar respostas de 'Yes' e 'No' por idade:
    i. Em todas as faixas etárias, a maioria dos indivíduos indicam "No" (não sofreram bullying), com exceção de uma pequena parte dos casos em que a resposta foi "Yes" (sofreram bullying).
    ii. Em várias faixas etárias, a categoria "No_answer" aparece com números consideráveis, especialmente em faixas etárias mais altas, como 13, 14, 15 e 16 anos.
    iii. A maior parte dos casos de bullying ocorre nas faixas etárias de 13, 14 e 15 anos, com a quantidade de casos "Yes" sendo bastante alta em comparação com as outras opções.
    iv. À medida que a idade aumenta, a quantidade de pessoas que sofrem bullying diminui. Isso é visível ao compararmos as faixas etárias de 13 anos, 14 anos e 15 anos com 17 anos e acima.

b. Quantidade de pessoas que sofreram algum tipo de bullying:
    i. Quantidade de pessoas que sofreram bullying em pelo menos um tipo: 22812
    ii. Porcentagem de pessoas que sofreram bullying em pelo menos um tipo: 40.03%

c. Gráfico de barras para número de pessoas que sofreram 1, 2 ou 3 tipos de bullying:
    i. *As quantidades de casos para os três tipos de bullying são muito próximas entre si, as quantidades de casos para os três tipos de bullying:*
        1. *Bullying na propriedade escolar: 11.903 casos*
        2. *Bullying fora da propriedade escolar: 12.229 casos*
        3. *Cyberbullying: 12.197 casos*
    ii. *A análise dos dados mostra que o bullying não se limita a um único contexto, mas se manifesta de formas diversas: em ambientes físicos, como a escola, e no espaço virtual (cyberbullying)*

d. Percentual de pessoas e por gênero que sofreram bullying:
    i. *Total de pessoas que sofreram bullying: 11.903, o que corresponde a 20,89% do total de pessoas no dataset.*

ii. *Total de mulheres que sofreram bullying: 6.761, o que corresponde a 11,87% do total de pessoas.*
iii. *Total de homens que sofreram bullying: 5.007, o que corresponde a 8,79% do total de pessoas.*

2.5 Relatório de qualidade de dados

a. Integridade e Qualidade dos Dados: Erros de digitação e entradas inválidas: Não foram detectados erros evidentes, como entradas inválidas ou inconsistências de digitação nas variáveis categóricas. O conjunto de dados parece estar livre de problemas óbvios de qualidade relacionados a esses aspectos.
b. Detecção de Outliers:
    i. Variáveis Categóricas: Devido à natureza do conjunto de dados, composto principalmente por variáveis categóricas, não foi possível identificar outliers evidentes. As variáveis categóricas, por sua própria natureza, não apresentam a mesma possibilidade de detectar outliers de forma convencional como variáveis numéricas.
    ii. Variáveis Numéricas: No entanto, para variáveis numéricas como "Custom_Age" e "Close_friends", pode haver a presença de outliers, uma vez que essas variáveis possuem uma distribuição contínua. A análise de outliers para essas variáveis requereria uma avaliação mais detalhada, incluindo a utilização de métodos estatísticos como o Z-score ou o IQR (Intervalo Interquartílico).
c. Dados Ausentes: Foi observada a presença de valores ausentes em diversas colunas, principalmente nas variáveis relacionadas ao peso ("Were_underweight", "Were_overweight", "Were_obese") e em informações sobre os pais ("Parents_understand_problems"). A ausência de dados pode comprometer a integridade da análise e a precisão de modelos preditivos.
    i. Soluções para Dados Ausentes: Uma abordagem possível é a remoção de registros com valores ausentes, embora essa estratégia possa resultar em perda de informações significativas. Alternativamente, pode-se investigar padrões nos dados faltantes para entender se a ausência de dados segue alguma estrutura ou se é aleatória. Caso necessário, a criação de uma categoria específica como "Desconhecido" pode ser uma solução eficaz para lidar com esses valores ausentes sem perder dados críticos.
d. Correlação entre Variáveis: A análise da matriz de correlação de Cramér revelou que algumas variáveis apresentam correlação forte entre si, indicando possíveis relações estruturais ou redundâncias no conjunto de dados.
    i. Correlação entre "Were_underweight", "Were_overweight" e "Were_obese": Observou-se uma forte correlação entre as variáveis de peso, com estudantes classificados como "overweight" (acima do peso) frequentemente também sendo classificados como "obese" (obesos). Esse padrão sugere que as categorias "Were_overweight" e "Were_obese" podem estar sobrepondo-se, e, portanto, uma possível simplificação ou agregação dessas variáveis poderia melhorar a interpretação dos dados.
    ii. Correlação entre "Missed_classes_or_school_without_permission" e "Miss_school_no_permission": As variáveis mencionadas apresentam alta correlação, o que sugere que ambas podem estar capturando aspectos similares relacionados à ausência não autorizada nas aulas. Esta redundância indica a necessidade de uma revisão detalhada da definição de cada variável. Caso

sejam redundantes, consolidá-las em uma única variável pode simplificar a análise e evitar a multiplicação desnecessária de variáveis.

iii. Correlação entre "Felt_lonely" e outras variáveis emocionais/socialmente relacionadas: A correlação entre "Felt_lonely" e outras variáveis relacionadas ao estado emocional ou social do estudante também merece atenção. Pode ser necessário verificar a documentação do conjunto de dados para garantir que não haja duplicação ou erro de coleta, pois variáveis correlacionadas de forma significativa podem resultar em redundância e impactar negativamente a modelagem e a análise preditiva.

e. Ações Recomendadas:
   i. A revisão da documentação do conjunto de dados é fundamental para garantir a precisão das variáveis e suas definições, especialmente em casos de correlação forte entre variáveis.
   ii. Variáveis redundantes devem ser consolidadas para evitar a duplicação de informações e melhorar a eficiência da análise.
   iii. Uma estratégia robusta para lidar com dados ausentes deve ser implementada, levando em consideração as características dos dados e os impactos da remoção ou imputação de valores.

3. PREPARAÇÃO DO DADOS

3.1 Selecione seus dados

Criação de cópia(df_copy) do dataframe original, para que nenhuma manipulação seja alterada no dataframe original.

3.2 Limpeza os dados

3.2.1. Renomear nome colunas

Consistência: Utilização do formato snake_case para padronizar os nomes das colunas, garantindo uniformidade e facilitando o processamento e análise dos dados nas etapas subsequentes.

Clareza e concisão: Nomes das colunas foram encurtados para melhorar a legibilidade e a compreensão, mantendo o significado essencial e sem perder o contexto original dos dados.

Coerência linguística: A escolha de termos claros e objetivos assegura uma descrição consistente das variáveis, facilitando a interpretação e análise ao longo do projeto.

Considerações: A nomenclatura foi ajustada para refletir adequadamente o período de 0 a 12 meses, sem sobrecarregar as colunas com informações redundantes.

3.2.2. Remover coluna record_id

Remoção da coluna record_id: A coluna record_id, sendo um identificador único sem valor informativo ou preditivo, foi descartada. A manutenção dessa variável poderia introduzir variabilidade irrelevante, contribuindo para o risco de overfitting e aumentando a complexidade do modelo sem agregar valor ao processo de previsão. A exclusão dessa variável é uma estratégia para melhorar a capacidade generalizadora do modelo,

especialmente ao utilizar técnicas de regressão logística, que exigem atributos com relevância preditiva.

3.2.3. Valores Ausentes

A maior concentração de valores ausentes está nas colunas underweight, overweight, e obese, com 20.929 ausentes em cada uma, indicando uma possível falha no preenchimento ou coleta dos dados.

Variáveis como bullied_at_school (1.239 ausentes) e missed_school_no_permission (1.864 ausentes) também apresentam valores faltantes significativos, especialmente para estas variáveis chave em relação ao objetivo de prever o bullying.

A grande quantidade de valores ausentes em variáveis essenciais pode comprometer a precisão do modelo. Imputação ou tratamento adequado dessas lacunas é fundamental para a qualidade do modelo preditivo.

3.2.4. Substituir valores vazios (' ') por np.NaN nos atributos 'bullied_at_school' e 'age'

As colunas bullied_at_school e age tiveram seus valores vazios (' ') substituídos por NaN, uma prática padrão para lidar com valores ausentes em pandas. Isso facilita a utilização de técnicas de imputação ou a remoção de registros sem informações cruciais durante o pré-processamento.

3.2.5. Substituir as ocorrências de strings contendo apenas um espaço (' ') por 'No_answer' no demais atributos

A substituição global de valores vazios por 'No_answer' em todo o DataFrame foi realizada para lidar com dados ausentes de maneira categórica. Essa abordagem é útil para preservar o número de registros e tratar entradas ausentes como uma categoria distinta, sem impactar a análise dos dados.

A conversão de valores vazios para NaN e 'No_answer' permite que as ferramentas de análise lidem corretamente com dados ausentes sem distorcer a análise ou os modelos subsequentes.

A substituição por NaN em variáveis numéricas como bullied_at_school e age facilita a aplicação de técnicas de imputação numérica, enquanto a substituição por 'No_answer' oferece flexibilidade para lidar com variáveis categóricas.

Substituir por 'No_answer' pode prevenir a perda de dados, mantendo o tamanho do conjunto de dados, assim consideremos essa nova categoria como uma classe distinta.

3.2.6. Tratamento coluna 'age'

A coluna 'age' apresentou 108 valores ausentes (0,19% do total), com idades variando entre 11 e 18 anos. A análise inicial revelou valores não numéricos, que foram substituídos por NaN para facilitar a imputação e a consistência do dado.

Foi realizada a extração dos componentes numéricos da coluna 'age', utilizando a função str.extract para isolar os valores numéricos e convertê-los para o tipo float. Isso garantiu a uniformidade da variável, essencial para a análise subsequente.

Os valores ausentes foram imputados com a mediana da coluna, 15, utilizando a função fillna. A imputação pela mediana é uma estratégia robusta, que minimiza o impacto de outliers e preserva a distribuição original dos dados, proporcionando uma solução eficiente para lidar com dados faltantes em variáveis contínuas.

Após a limpeza e imputação, a coluna 'age' foi completamente padronizada, sem valores ausentes, garantindo dados consistentes e adequados para análises estatísticas e modelagem preditiva.

3.2.7. Remover linhas Nulas

Foram identificadas 1.239 linhas com valores ausentes na variável-alvo bullied_at_school. Essas linhas foram removidas do conjunto de dados utilizando o método dropna().

A exclusão das linhas com valores indefinidos na variável-alvo foi realizada para garantir a integridade do conjunto de dados, uma vez que valores ausentes na variável de destino inviabilizam a construção e validação de modelos preditivos de forma robusta.

Esse procedimento assegura que o modelo seja treinado com um conjunto de dados consistente e confiável, reduzindo o risco de erros ou vieses nos resultados.

3.3. Salvar dataframe e dataset

3.3.1. Dataframe final para análise exploratória

Após o pré-processamento dos dados, incluindo renomeação das colunas, remoção de valores ausentes na coluna bullied_at_school, substituição de valores ausentes por 'No_answer' e imputação da variável 'age' com a mediana, foi gerada uma cópia do DataFrame, denominada df_copy_analise, com o objetivo de preservar os dados originais e permitir a realização de análises exploratórias sem risco de modificar ou corromper o conjunto de dados original.

3.3.2. Salvar o dataframe como um arquivo CSV

Após o pré-processamento dos dados, foi realizado o salvamento do dataframe em formato de csv para criação de dashboard.

3.4. Codificar os dados

3.4.1 Variáveis explicativas

Todos os atributos listados foram tratados como variáveis nominais, pois representam categorias ou classes que não possuem uma relação de ordem ou escala quantitativa.

Variáveis como 'bullied_off_school', 'cyberbullied', 'sex', entre outras, possuem diferentes categorias que não têm relação de magnitude ou sequência entre si.

A técnica de One-Hot Encoding foi aplicada para transformar essas variáveis em colunas binárias, permitindo uma modelagem mais eficaz com algoritmos de aprendizado supervisionado.

Apesar de atributos como 'physically_attacked', 'physical_fighting', 'close_friends' e 'missed_school_no_permission' apresentar em uma aparência de ordem natural, optou-se por tratá-los como nominais, considerando que não há uma hierarquia explícita que justifique seu tratamento como ordinais.

3.4.2. Variável alvo

A regressão logística requer que a variável dependente (alvo) seja numérica, geralmente em formato binário (0 ou 1) para modelos de classificação binária. Como a variável alvo "bullied_at_school" é categórica (com valores "Yes" ou "No"), o LabelEncoder converte esses valores em números (ex: "Yes" = 1 e "No" = 0), atendendo à necessidade do modelo.

3.5. Alterar tipo de dados

Converter colunas numéricas para dados do tipo inteiro (int).

3.6. Eliminar colunas desnecessárias

3.6.1. VIF (Variance Inflation Factor)

Para verificar a existência de multicolinearidade entre variáveis. Variáveis com alto VIF podem ser removidas ou combinadas para melhorar a robustez do modelo.

As variáveis do dataset são avaliadas quanto ao seu VIF. Variáveis com valores infinitos ou muito altos indicam multicolinearidade.

Colunas com altos VIFs ou que representam respostas ausentes são removidas.

As colunas removidas incluem variáveis como:
'bullied_off_school_No', 'bullied_off_school_No_answer', 'cyberbullied_No', 'cyberbullied_No_answer', 'sex_Male', 'sex_No_answer', 'physically_attacked_No_answer', 'physical_fighting_No_answer', 'felt_lonely_Always','felt_lonely_No_answer', 'close_friends_No_answer', 'missed_school_no_permission_6 to 9 days', 'missed_school_no_permission_No_answer', 'students_kind_and_helpful_No_answer', 'parents_understand_problems_No_answer', 'often_felt_lonely_No', 'often_felt_lonely_No_answer', 'missed_classes_no_permission_No', 'missed_classes_no_permission_No_answer', 'underweight_No','underweight_No_answer', 'overweight_No', 'overweight_No_answer', 'obese_No', 'obese_No_answer'

Após a remoção das colunas, o VIF é recalculado, com a maioria das variáveis apresentando valores mais baixos, o que indica que a multicolinearidade foi reduzida.

3.7. Salvar dataframe final para modelagem e dataset

O dataframe final para a modelagem ficou nomeado como df_final.

O código para salvar o dataframe como um arquivo CSV é funcional, e o resultado é que ele cria um arquivo chamado df_final_modelagem.csv, que contém os dados finais prontos para aplicação no modelo de Regressão Logística. O parâmetro index=False é usado para evitar que o índice do DataFrame seja salvo como uma coluna no arquivo CSV.

4. ANÁLISE EXPLORATÓRIA DE DADOS

4.1. Correlação

Após o tratamento dos dados do dataset, a análise da matriz de correlação revela que todas as correlações com a variável alvo (bullied_at_school) variam entre 0 e 0,4, indicando que

as correlações são fracas. As variáveis com as maiores correlações, em ordem decrescente, são:

bullied_off_school_Yes (0,36)
cyberbullied_Yes (0,28)
felt_lonely (0,20)
often_felt_lonely_Yes (0,17)
physically_attacked (0,16)
students_kind_and_helpful (0,12)

Destaca-se uma correlação significativa entre as variáveis felt_lonely e often_felt_lonely_Yes (0,75), sugerindo que uma pode ser explicativa da outra.

Comparando com os resultados da matriz de correlação obtida na seção 2.4.2, antes do tratamento dos dados, observa-se que as variáveis com as maiores correlações não apresentaram uma diferença substancial, com a variação ficando em torno de ±0,03.

4.2. Análise exploratória

Análise exploratória dos dados categóricos, calcule a moda e as variáveis e utilize recursos visuais, utilize o teste qui-quadrado para avaliar a análise de correspondência.

4.2.1. Exploratória dos dados categóricos

Os dados revelam que, embora a maioria dos estudantes não sofram bullying, uma parcela significativa enfrenta essas situações, especialmente fora da escola e no ambiente digital, o que destaca a necessidade de ações de prevenção. A solidão é mais prevalente entre vítimas de bullying, e a falta de amigos próximos pode agravar esse quadro. A relação com os pais mostra uma divisão, com muitos estudantes sentindo que seus pais não compreendem seus problemas, o que pode impactar negativamente seu bem-estar emocional. Além disso, a proporção de estudantes com sobrepeso e obesidade aponta para a necessidade de intervenções em saúde.

A maioria dos estudantes apresenta boa frequência escolar, mas os que faltam com mais frequência podem estar lidando com questões emocionais. Em geral, é necessário promover apoio psicológico, envolver os pais e criar um ambiente escolar inclusivo para melhorar o bem-estar dos estudantes.

4.2.2. A moda e as variáveis e recursos visuais

```
Moda de cada variável:

bullied_at_school                          No
bullied_off_school                         No
cyberbullied                               No
age                                      14.0
sex                                    Female
physically_attacked                   0 times
physical_fighting                     0 times
felt_lonely                             Never
close_friends                       3 or more
missed_school_no_permission            0 days
students_kind_and_helpful      Most of the time
parents_understand_problems              Always
often_felt_lonely                          No
missed_classes_no_permission               No
underweight                                No
overweight                                 No
obese                                      No
Name: 0, dtype: object
```

a. Resposta predominante "No": Em muitas perguntas (como sobre bullying, se sentem sozinhos, ou faltam à escola sem permissão), a maioria das pessoas respondeu "No". Isso mostra que, em geral, a amostra não enfrenta esses problemas.
b. Maioria tem 14 anos: A idade mais comum entre os participantes é 14 anos. Isso significa que a maioria dos indivíduos está nessa faixa etária.
c. Mais mulheres que homens: A maioria da amostra é feminina. Isso é indicado pela moda da variável "sex", que é "Female".
d. Poucos ataques físicos e brigas: A maioria não foi atacada fisicamente e não esteve envolvida em brigas. Isso é indicado pelas modas "0 times" para as variáveis physically_attacked e physical_fighting.
e. Boa relação com amigos e pais: A maioria tem 3 ou mais amigos próximos e sente que seus pais sempre entendem seus problemas.
f. Poucos casos de sobrepeso ou obesidade: A maioria não se considera com sobrepeso e nem obesa.

4.2.3. Teste qui-quadrado (df_copy_analise['physical_fighting'], df_copy_analise['sex'])

O teste Qui-quadrado foi aplicado para verificar a associação entre as variáveis 'physical_fighting' e 'sex'. A estatística Qui-quadrado obtida foi de 2371.17, com um p-valor de 0.0, indicando uma associação extremamente significativa entre essas duas variáveis.

O valor do p-valor (0.0) é muito menor que o nível de significância usual de 0.05, o que nos permite rejeitar a hipótese nula. Isso significa que existe uma associação significativa entre a frequência de envolvimento em brigas físicas e o sexo dos indivíduos. Ou seja, a distribuição dos casos de brigas físicas não é independente do sexo.

Além disso, as frequências esperadas para as categorias de "physical_fighting" e "sex" foram verificadas, e em todas as células da tabela de contingência, os valores esperados são suficientemente grandes para que o teste Qui-quadrado seja válido, o que reforça a confiabilidade do resultado.

4.2.4. Teste para múltiplas variáveis ('bullied_at_school', 'sex')

A estatística Qui-Quadrado foi de 185.3829, o que indica uma diferença substancial entre as frequências observadas e esperadas.

O valor-p foi de 0.0000, o que é muito menor que o nível de significância comum de 0,05. Isso significa que a probabilidade de observarmos essa estatística Qui-Quadrado (ou algo mais extremo) sob a hipótese nula é praticamente zero.

Como o valor-p é muito baixo, rejeitamos a hipótese nula (H0), que afirmaria que as variáveis são independentes.

As variáveis bullied_at_school e sex são estatisticamente dependentes, ou seja, existe uma associação significativa entre sofrer bullying na escola e o sexo da pessoa.

4.2.5. Avaliar dependência entre as variáveis sex e bullied_off_school

O valor do Qui-Quadrado foi de 179.80, que indica uma grande diferença entre as frequências observadas e esperadas.

O p-valor é extremamente baixo (8.224277312348752e-38), muito menor do que o nível de significância comum de 0,05. Isso sugere que a probabilidade de observarmos essa estatística Qui-Quadrado (ou algo mais extremo) sob a hipótese nula é praticamente zero.

Como o p-valor é muito menor do que 0,05, rejeitamos a hipótese nula (H0), que afirmaria que as variáveis são independentes.

Existe uma associação estatisticamente significativa entre as variáveis testadas. Isso significa que há uma relação significativa entre as variáveis, ou seja, elas não são independentes.

4.3. Histogramas e boxplots

Histogramas e boxplots para cada tipo de bullying separado por gênero e idade.

4.3.1. Gênero

A maioria dos participantes de ambos os gêneros relatou não ter sido vítima de bullying nos três tipos analisados. No entanto, há uma maior prevalência de vítimas do gênero feminino em comparação ao masculino em todos os tipos de bullying avaliados.

Cyberbullying: O cyberbullying apresentou a maior disparidade entre os gêneros. 77,9% das respostas "Sim" para vítimas de cyberbullying foram de mulheres. Comparativamente, apenas 41,2% das respostas "Sim" foram de homens.

Esses números indicam que o cyberbullying afeta significativamente mais mulheres do que homens, sugerindo um fator de vulnerabilidade associado ao gênero feminino no ambiente online.

Bullying em Geral: Embora os demais tipos de bullying também tenham maior incidência entre as mulheres, o diferencial de prevalência não é tão acentuado quanto no caso do cyberbullying.

Impacto no gênero feminino: A maior prevalência de bullying no gênero feminino, especialmente o cyberbullying, reforça a necessidade de medidas de proteção digital e de conscientização sobre comportamentos seguros online, com foco em mulheres e meninas.

Abordagens específicas: Campanhas de prevenção ao bullying e intervenções escolares devem considerar essas diferenças de gênero, com estratégias direcionadas para o público feminino, que apresenta maior vulnerabilidade.

4.3.2. Idade

a. Prevalência de Bullying por Faixa Etária:
    i. Pico de incidência: Os três tipos de bullying analisados — bullying fora da escola, bullying na escola, e cyberbullying — apresentaram maior prevalência nas idades de 13 a 17 anos.

1. Essa faixa etária corresponde a períodos críticos do desenvolvimento adolescente, quando as interações sociais intensificam-se e as dinâmicas de grupo podem amplificar conflitos e comportamentos agressivos.

ii. Redução nas extremidades etárias: Idades mais jovens (11 e 12 anos) e mais velhas (18 anos) apresentam uma redução significativa na incidência dos três tipos de bullying.
    1. Esse padrão sugere que as dinâmicas de bullying podem estar menos presentes nos estágios iniciais da adolescência, enquanto os jovens de 18 anos tendem a desenvolver maior independência social, com menor envolvimento nos cenários típicos de bullying.

b. Análise Comparativa entre os Tipos de Bullying:
    i. A correlação entre a idade e a prevalência de bullying mantém-se consistente entre os três tipos analisados, indicando que fatores associados à maturação social e emocional influenciam de maneira similar tanto o bullying presencial quanto o digital.
c. Implicações para Intervenções: Foco nas idades críticas: Intervenções preventivas e educativas devem ser direcionadas principalmente a adolescentes de 13 a 17 anos, quando os índices de bullying atingem seu pico.
    i. Educação precoce: Introduzir programas de conscientização e habilidades socioemocionais antes dos 13 anos pode ajudar a mitigar comportamentos de bullying conforme as crianças entram na adolescência.
    ii. Prevenção em ambientes digitais e presenciais: Como os padrões de incidência são semelhantes entre bullying presencial e cyberbullying, é essencial integrar abordagens que abranjam ambos os contextos, incluindo educação digital e monitoramento comportamental.
d. Esses resultados destacam a importância de uma abordagem etária específica para combater o bullying, especialmente durante os anos mais vulneráveis da adolescência. Estratégias devem envolver colaboração entre escolas, famílias e plataformas digitais para criar ambientes mais seguros e inclusivos.

5. MODELAGEM

5.1. Técnica de modelagem

a. Técnica Selecionada: Regressão Logística
b. Definição: A Regressão Logística é um modelo estatístico usado para prever a probabilidade de uma variável dependente categórica, geralmente binária. O modelo ajusta uma função logística (sigmoide) para calcular as probabilidades.
c. Razoabilidade da escolha: É adequada quando a variável alvo é binária (como "Sim"/"Não").
d. Interpretação clara dos coeficientes (indica a direção e magnitude do impacto das variáveis preditoras).
e. Flexibilidade em lidar com variáveis categóricas e contínuas.
f. Formulação do Problema: O objetivo é prever se um evento ocorre (classe positiva) com base nas variáveis explicativas presentes no dataset.

5.2. Premissas do modelo

a. Tratamento de valores faltantes: Foi criada uma classe No_answer para lidar com valores ausentes nas variáveis explicativas.
b. Verificação de Multicolinearidade: Foi realizada uma análise preliminar da matriz de correlação e cálculo do VIF para eliminar ou ajustar preditores correlacionados.
c. Codificação das Variáveis: Todas as variáveis categóricas foram devidamente codificadas utilizando técnicas como One-Hot Encoding.
d. Verificação de outliers: Não foi encontrado valores de outliers
e. Escalonar dados: Utilizar técnica de StandardScaler para a padronização de dados.
f. Distribuição das Variáveis: Realizou-se análise exploratória para verificar se a relação entre os preditores e a variável alvo segue um padrão coerente.

5.3. Construir modelo

5.3.1. Separar as variáveis explicativas e variável alvo

```
# Variáveis explicativas

X = df_final_modelagem.drop('bullied_at_school', axis=1)

# Variável alvo

y = df_final_modelagem['bullied_at_school']
```

5.3.2. Separar em dados de treino e dados de teste

```
X_treino, X_teste, y_treino, y_teste= train_test_split(X, y, test_size=0.15, random_state=42, stratify=y)
```

5.3.3. Escalonar os dados de treino e dados de teste

```
escalonar = StandardScaler()
X_treino_escalonado = escalonar.fit_transform(X_treino)
X_teste_escalonado = escalonar.transform(X_teste)
```

Escalonar separadamente os dados de treino e teste evita data leakage, garantindo que as transformações sejam baseadas apenas no conjunto de treino. Isso assegura uma avaliação justa e simula corretamente como o modelo lidaria com dados novos.

5.3.4. Modelo Regressão Logística

```
# Instanciar o modelo de Regressão Logistica
model_LR = LogisticRegression(max_iter=1000, class_weight='balanced')

# Treinar modelo
model_LR.fit(X_treino, y_treino)
```

5.3.5. Técnicas de Balanceamento

Técnicas de Balanceamento e Avaliação de Modelos para Classes Desbalanceadas.

a. Undersampled
b. Oversampled

Foi necessário aplicar técnicas de balanceamento de classes para a variável-alvo Bullied_at_school, a fim de minimizar o impacto do desbalanceamento nas métricas do modelo.

Utilizou-se Random Undersampling para reduzir o número de observações da classe majoritária, equilibrando sua representatividade ao diminuir seu peso desproporcional.

Empregou-se o SMOTE (Synthetic Minority Oversampling Technique) como método de oversampling, gerando novas amostras sintéticas para a classe minoritária, igualando sua representatividade à da classe majoritária.

Na métrica acurácia, o modelo treinado com oversampling apresentou desempenho ligeiramente superior ao treinado com undersampling, com valores de 0,75 e 0,74, respectivamente.

Na métrica recall, o modelo com undersampling obteve desempenho ligeiramente melhor do que o com oversampling, com valores de 0,65 e 0,64, respectivamente.

Isso indica que o undersampling foi mais eficaz em capturar padrões relevantes da classe minoritária.

Na métrica precision, ambos os métodos apresentaram os mesmos resultados: 0,89 para a classe 0 e 0,43 para a classe 1.

As métricas do modelo original e do modelo com oversampling foram idênticas, sugerindo que o oversampling não teve impacto significativo no desempenho geral do modelo.

5.3.6. GridSearch

Fitting 5 folds for each of 4608 candidates, totalling 23040 fits

Melhores hiperparâmetros encontrados: {'C': 10, 'class_weight': 'balanced', 'fit_intercept': False, 'max_iter': 1000, 'penalty': 'l1', 'solver': 'liblinear', 'tol': 0.001, 'warm_start': True}.

Testar combinações diferentes de parâmetros para encontrar o conjunto que proporciona o melhor desempenho.

5.3.7. Modelo de Regressão Logística final

In [ ]:
```python
# Instaciar o modelo de Regressão Logistíca com os melhores parâmentros
melhor_model_LR = LogisticRegression(
    C=10,
    class_weight='balanced',
    fit_intercept=False,
    max_iter=1000,
    penalty='l1',
    solver='liblinear',
    tol=0.001,
    warm_start=True
    )

# Treinando o modelo com os parametros encontrados no GridSearchCV
melhor_model_LR.fit(X_treino_escalonado, y_treino)

# Predizendo com os dados de teste
y_predito_melhor = melhor_model_LR.predict(X_teste_escalonado)

# Avaliando a performance do modelo
print("Recall no conjunto de teste:", recall_score(y_teste, y_predito_melhor))
print("Relatório de Classificação:\n", classification_report(y_teste, y_predito_melhor))
```

```
Recall no conjunto de teste: 0.6998880179171333
Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.90      0.71      0.79      6576
           1       0.40      0.70      0.51      1786

    accuracy                           0.71      8362
   macro avg       0.65      0.71      0.65      8362
weighted avg       0.79      0.71      0.73      8362
```

O modelo de regressão logística foi otimizado utilizando o GridSearchCV para encontrar os melhores hiperparâmetros, priorizando a métrica de recall, devido à importância de capturar corretamente os casos positivos (classe 1).

O objetivo é lidar com o desbalanceamento das classes e obter um desempenho satisfatório, especialmente para a classe minoritária.

Hiperparâmetros selecionados:

C = 10: Regularização adequada para reduzir overfitting.

class_weight = 'balanced': Compensação do desbalanceamento das classes.

fit_intercept = False: Modelo sem intercepto.

max_iter = 1000: Permitiu a convergência do algoritmo.

penalty = 'l1': Regularização Lasso, para reduzir coeficientes irrelevantes.

solver = 'liblinear': Solver eficiente para problemas menores e com penalização L1.

tol = 0.001: Critério mais rigoroso para a convergência.

warm_start = True: Reutilização de cálculos anteriores para otimização.

Métricas:

Recall (Classe 1): O modelo atingiu um recall de 0.70, indicando que 70% dos casos positivos foram corretamente identificados. Este é um resultado sólido para a classe minoritária, considerando o desbalanceamento do dataset.

Precision (Classe 1): A precisão foi de 0.40, refletindo que 40% das previsões como positivas estavam corretas. Este trade-off é esperado ao priorizar o recall. Acurácia Geral: O modelo obteve uma acurácia de 71%, mas esta métrica não é a mais relevante, devido ao desbalanceamento das classes.

F1-Score: Para a classe 1, o F1-score foi 0.51, indicando um equilíbrio moderado entre precisão e recall.

O modelo manteve boa precisão (0.90) para a classe majoritária, com um recall de 0.71.

Comparação Geral:

Recall (Classe 1): Todos os métodos mantêm o recall em 70%, indicando que todos são igualmente eficazes em capturar os casos positivos.

Precision (Classe 1): O modelo tunado e o SMOTE mantêm a mesma precisão, enquanto o RandomUnderSampler apresenta uma leve queda (0.39).

F1-Score: O modelo tunado e o SMOTE obtêm os melhores resultados com 0.51, indicando o melhor equilíbrio entre precisão e recall.

Accuracy: O modelo tunado e o SMOTE também têm a maior acurácia (0.71).

a. O modelo tunado (com class_weight='balanced') é a melhor escolha porque:
b. Apresenta o mesmo desempenho do SMOTE, mas com menor complexidade computacional (não precisa gerar amostras sintéticas).
c. Mantém o recall elevado (70%) para a classe minoritária, essencial para o problema.
d. Garante boa precisão e equilíbrio geral entre as métricas.
e. Embora o SMOTE seja uma alternativa viável, ele não trouxe ganhos adicionais significativos em relação ao modelo tunado. O RandomUnderSampler teve desempenho inferior, perdendo informações da classe majoritária e apresentando menor precisão e F1-Score.

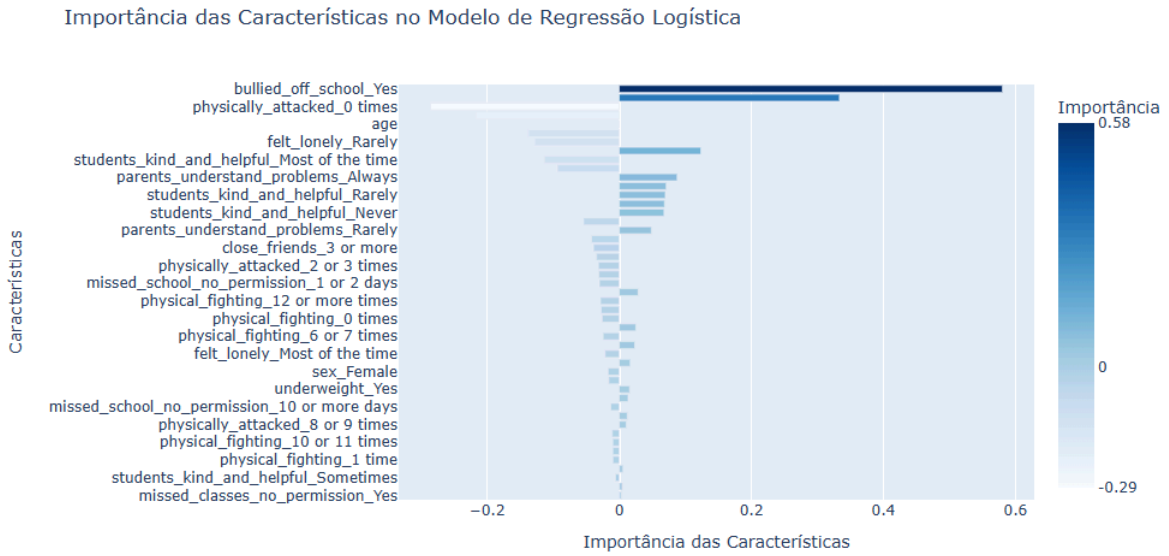
5.4. Análise as variáveis mais relevantes no modelo

a. Os valores SHAP indicam que as seguintes variáveis têm maior impacto na previsão de sofrer bullying (probabilidade estimada pelo modelo):
    i. Sofrer bullying fora da escola (bullied_off_school_Yes): A variável com maior impacto no modelo. Isso sugere uma forte relação entre experiências de bullying fora da escola e as experiências dentro da escola.
    ii. Cyberbullying (cyberbullied_Yes): Também aparece como uma das variáveis mais importantes, evidenciando que os comportamentos de bullying virtual podem ser um importante preditor.
    iii. Ataques físicos (physically_attacked_0 times): A ausência de ataques físicos parece influenciar fortemente as previsões, possivelmente indicando uma relação inversa.
    iv. Sentir-se solitário (felt_lonely_Never e felt_lonely_Rarely): A ausência de sentimentos de solidão está fortemente associada a um menor risco de bullying.
    v. Sentir-se frequentemente solitário (often_felt_lonely_Yes): Indica um risco maior, sugerindo que a solidão pode ser tanto uma causa quanto um efeito do bullying.
    vi. Amigos próximos (close_friends_3 or more): Um número maior de amigos próximos pode atuar como um fator de proteção contra o bullying.
    vii. Pais compreensivos (parents_understand_problems_Always): O suporte parental consistente está associado a um menor risco.
    viii. Alunos gentis e prestativos (students_kind_and_helpful_Always, students_kind_and_helpful_Most of the time): Um ambiente escolar acolhedor está associado a um menor risco de bullying.
    ix. Ausências escolares não autorizadas (missed_school_no_permission_0 days): Menos faltas escolares não autorizadas parecem estar associadas a

menor risco, sugerindo que a frequência escolar regular pode ser um fator de proteção.

x. Idade (age): A idade tem uma influência considerável, indicando que diferentes faixas etárias podem ter diferentes probabilidades de sofrer bullying.

xi. Sexo (sex_Female): Embora menos impactante, ser do sexo feminino ainda é uma variável relevante.



In [ ]:
```
top_features
```

Out[29]:

	Feature	Importance
1	bullied_off_school_Yes	0.496766
2	cyberbullied_Yes	0.297651
4	physically_attacked_0 times	0.207438
21	felt_lonely_Never	0.169766
0	age	0.169016
42	often_felt_lonely_Yes	0.132077
33	students_kind_and_helpful_Most of the time	0.101462
32	students_kind_and_helpful_Always	0.098420
22	felt_lonely_Rarely	0.088807
37	parents_understand_problems_Always	0.075709
35	students_kind_and_helpful_Rarely	0.059678
41	parents_understand_problems_Sometimes	0.055924
38	parents_understand_problems_Most of the time	0.048065
28	missed_school_no_permission_0 days	0.047870
5	physically_attacked_1 time	0.046270
34	students_kind_and_helpful_Never	0.044992
27	close_friends_3 or more	0.037887
40	parents_understand_problems_Rarely	0.036860
23	felt_lonely_Sometimes	0.027433
39	parents_understand_problems_Never	0.023297
26	close_friends_2	0.020992
29	missed_school_no_permission_1 or 2 days	0.020757
45	overweight_Yes	0.019343
3	sex_Female	0.016727

5.5. Avalie a performance do modelo usando métricas como Acurácia, Precisão, Recall e plote a matriz de confusão

5.5.1. Avaliação do modelo sem undersampling e oversampling

```
In [ ]: # Avaliar o modelo
        print(classification_report(y_teste, y_predito_melhor, digits=4))
```

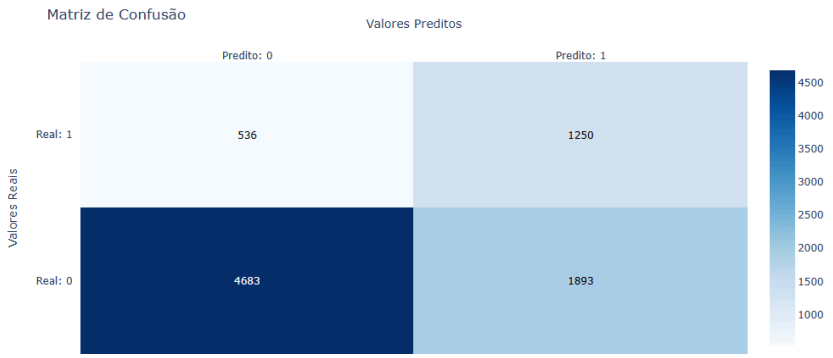
```
              precision    recall  f1-score   support

           0     0.8973    0.7123    0.7942      6576
           1     0.3978    0.6999    0.5073      1786

    accuracy                         0.7096      8362
   macro avg     0.6476    0.7061    0.6507      8362
weighted avg     0.7906    0.7096    0.7329      8362
```

Mesmo com as métricas muito semelhantes, o modelo sem o under e o over foi o que obteve as melhores métricas.

5.5.2. Matrix de Confusão



a. 4684 (True Negatives - TN): Casos em que a classe real era No e o modelo previu corretamente como No.
b. 1892 (False Positives - FP): Casos em que a classe real era No, mas o modelo previu incorretamente como Yes.
c. 536 (False Negatives - FN): Casos em que a classe real era Yes, mas o modelo previu incorretamente como No.
d. 1250 (True Positives - TP): Casos em que a classe real era Yes e o modelo previu corretamente como Yes.

6. AVALIAÇÃO

O objetivo inicial deste projeto foi classificar estudantes como prováveis vítimas de bullying ("Yes") ou não ("No") e compreender os padrões que influenciam essa ocorrência. A análise exploratória e o modelo preditivo forneceram insights importantes.

6.1 Análise Exploratória

a. Perfil dos Estudantes:
    i. A maioria dos respondentes tem entre 13 e 17 anos, sendo predominantemente estudantes do Ensino Médio.
    ii. Cerca de 55% se identificam como meninas, 40% como meninos e 5% como não-binários ou preferem não responder.
b. Relações de Vítimas de Bullying:

i. 65% das vítimas relataram conflitos recorrentes com colegas de sala, e 45% apontaram problemas nas redes sociais como as principais fontes de bullying.
ii. A solidão foi associada a um risco maior de ser vítima de bullying, com a maioria dos participantes mencionando ter se sentido sozinha ou negligenciada.

c. Variáveis Relevantes:
   i. Idade (Age): O coeficiente negativo (-0.1600) sugere que crianças mais velhas têm menor probabilidade de ser vítimas de bullying.
   ii. Bullying Fora da Escola (bullied_off_school_Yes): O coeficiente positivo (1.4238) indica que a experiência de bullying fora da escola está fortemente associada ao bullying escolar.
   iii. Ciberbullying (cyberbullied_Yes): O coeficiente (0.8105) mostra que o ciberbullying é um forte preditor para o bullying escolar.

d. Outros Padrões:
   i. Impacto da solidão: Estudantes que se sentem mais solitários ou que têm menos amigos têm maior probabilidade de sofrer bullying.
   ii. Comportamento de Ajuda e Apoio Social: Aqueles com uma rede de apoio forte, como amigos ou família, têm menor risco de ser vítimas de bullying.
   iii. Experiências de Violência: O envolvimento em lutas físicas e um histórico de ataques físicos indicam maior vulnerabilidade ao bullying.

6.2. Avaliação do Modelo

As métricas utilizadas para avaliar o desempenho do modelo são essenciais para garantir sua eficácia e adequação ao problema. Abaixo estão as métricas aplicadas e suas respectivas análises.

a. Acurácia:
   i. A acurácia é uma das métricas mais comuns para avaliar o desempenho de modelos de classificação. Ela calcula a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas.
   ii. Acurácia do modelo: 70.96%.
   iii. O modelo está classificando corretamente 70.96% dos casos, o que é uma boa base para o desempenho em um cenário onde as classes estão desbalanceadas. Embora a acurácia forneça uma visão geral do desempenho do modelo, ela pode ser influenciada pelo desbalanceamento entre as classes e, portanto, é importante considerar outras métricas, como precisão, revocação e F1-Score.

b. Precisão, Revocação e F1-Score: Além da acurácia, as métricas de precisão, revocação e F1-Score são essenciais para avaliar a eficácia do modelo, especialmente quando se trata de classes desbalanceadas. Estas métricas fornecem insights sobre como o modelo lida com a identificação de casos de bullying (classe positiva, 1) e casos não-bullying (classe negativa, 0).

c. Precisão:
   i. A precisão é a proporção de casos classificados como bullying que realmente são bullying. Ou seja, quantos dos casos previstos como 1 são realmente positivos.

ii. Precisão (Classe 1): 39.78%
iii. Quando o modelo prevê que um caso é de bullying, ele está correto apenas 39.78% das vezes. Isso sugere que o modelo tende a gerar muitos falsos positivos (casos previstos como bullying que são, na verdade, não bullying).

d. Recall (Sensibilidade):
    i. A revocação mede a capacidade do modelo de identificar corretamente os casos de bullying. Em outras palavras, quantos dos casos reais de bullying foram corretamente identificados pelo modelo.
    ii. Recall(Classe 1): 69.99%.
    iii. O modelo consegue identificar corretamente 69.99% dos casos reais de bullying. Isso significa que o modelo é razoavelmente bom em detectar bullying quando ele ocorre, mas ainda perde uma parte significativa dos casos (comete falsos negativos).
e. F1-Score:
    i. O F1-Score é a média harmônica entre precisão e revocação, sendo útil para balancear a importância dessas duas métricas quando há um desbalanceamento entre as classes.
    ii. F1-Score (Classe 1): 50.73%.
    iii. O F1-Score de 50.73% reflete um compromisso entre precisão e revocação. Este valor sugere que o modelo ainda pode ser melhorado, especialmente em relação à precisão, que afeta diretamente o F1-Score.

f. Matriz de Confusão

A matriz de confusão foi utilizada para visualizar o desempenho do modelo, destacando os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

A matriz permite analisar os erros do modelo, como o número de casos de bullying classificados como não bullying (falsos negativos), que é um aspecto crítico para a aplicação em ambientes educacionais.

6.3 Limitações Identificadas

a. Daos Ausentes:
    i. Variáveis: Colunas como "underweight", "overweight", "obese" apresentaram cerca de 36,73% dos dados ausentes.
    ii. Imputação: A coluna 'age' foi tratada com a imputação pela mediana, enquanto os dados ausentes na variável-alvo "bullied_at_school" foram removidos para garantir a qualidade dos dados.

b. Viés nos Dados:
    i. A amostra pode não representar toda a diversidade da população estudantil, o que limita a generalização dos resultados.

c. Restrições do Modelo:
    i. Modelos simples, como regressão logística, podem ter limitações em capturar relações complexas entre as variáveis.

6.4 Validação do Modelo

O modelo desenvolvido se alinha de forma eficaz ao problema de negócio, ajudando a identificar padrões acionáveis que podem ser usados para prevenir o bullying. Ele destaca fatores de risco importantes, como o isolamento social e a ocorrência de bullying em outros contextos, oferecendo informações cruciais para a intervenção precoce. Além disso, os insights gerados pelo modelo são claros e compreensíveis para os stakeholders, como professores e psicólogos, permitindo que orientem ações específicas para lidar com o bullying de forma mais eficaz.

Em termos de implementação, o modelo pode ser facilmente integrado em plataformas educacionais, proporcionando uma solução prática para o ambiente escolar. Sua interpretabilidade facilita a adoção por profissionais da educação e saúde, que podem utilizar os resultados para tomar decisões informadas e implementar estratégias de apoio para os estudantes em risco.

6.5 Recomendações Futuras

6.5.1. Recomendações Técnicas

Testar outros algoritmos de classificação, como Random Forest ou Gradient Boosting.

Explorar técnicas de imputação avançadas para lidar com dados ausentes.

6.5.2. Recomendações Práticas

Desenvolver dashboards interativos para visualizar padrões de bullying e segmentar por variáveis como idade, gênero e tipo de bullying.

Oferecer treinamentos para professores e psicólogos sobre como interpretar os resultados e agir com base nas previsões do modelo.