INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CAMPUS CAMPINAS

BIANCA VITÓRIA DOS SANTOS VIANA

PREDIÇÃO DE NOTAS DE ALUNOS DO ENSINO MÉDIO A PARTIR DE INDICADORES ESCOLARES E DE *BULLYING* USANDO APRENDIZADO DE MÁQUINA

CAMPINAS
2023

BIANCA VITÓRIA DOS SANTOS VIANA

PREDIÇÃO DE NOTAS DE ALUNOS DO ENSINO MÉDIO A PARTIR DE INDICADORES ESCOLARES E DE *BULLYING* USANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.

Orientador: Prof. Dr. Samuel Botter Martins

CAMPINAS

2023

Ficha catalográfica
Instituto Federal de São Paulo – Câmpus Campinas
Biblioteca
Rosangela Gomes – CRB 8/8461

Viana, Bianca Vitória dos Santos
V614p Predição de notas de alunos do ensino médio a partir de indicadores escolares e de bullying usando aprendizado de máquina/ Bianca Vitória dos Santos Viana. – Campinas, SP: [s.n.], 2023.
53 f. : il.

Orientador: Samuel Botter Martins
Trabalho de Conclusão de Curso (graduação) – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, 2023.

1. Bullying. 2. Desempenho escolar. 3. Aprendizado de máquina. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas. II. Título.

ATA N.º 4/2023 - TEINFO-CMP/DAE-CMP/DRG/CMP/IFSP

Ata de Defesa de Trabalho de Conclusão de Curso - Graduação

Na presente data, realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso intitulado **PREDIÇÃO DE NOTAS DE ALUNOS DO ENSINO MÉDIO A PARTIR DE INDICADORES ESCOLARES E DE BULLYING USANDO APRENDIZADO DE MÁQUINA** , apresentado(a) pelo(a) aluno(a) **BIANCA VITÓRIA DOS SANTOS VIANA (CP300547X)** do Curso **SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS** (Câmpus Campinas). Os trabalhos foram iniciados às **15h00** pelo(a) Professor(a) presidente da banca examinadora, constituída pelos seguintes membros:

Membros	Instituição	Presença (Sim/Não)
SAMUEL BOTTER MARTINS (Presidente/Orientador)	IFSP-CMP	SIM
DANILO DOURADINHO FERNANDES (Examinador 1)	IFSP-CMP	SIM
RICARDO BARZ SOVAT (Examinador 2)	IFSP-CMP	SIM

Observações:

A banca examinadora, tendo terminado a apresentação do conteúdo da monografia, passou à arguição do(a) candidato(a). Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo(a) aluno(a), tendo sido atribuído o seguinte resultado:

[**X**] Aprovado(a) [ ] Reprovado(a)

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu lavrei a presente ata que assino em nome dos demais membros da banca examinadora.

Câmpus Campinas, 23 de maio de 2023

Documento assinado eletronicamente por:

- **Samuel Botter Martins**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 23/05/2023 16:09:35.
- **Ricardo Barz Sovat**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 23/05/2023 16:10:11.
- **Danilo Douradinho Fernandes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 23/05/2023 18:02:03.

Este documento foi emitido pelo SUAP em 23/05/2023. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 554171
Código de Autenticação: d9b264537b

Dedico este trabalho a toda minha família,
e a todas as mulheres que, assim como eu,
já duvidaram de si mesmas,
para lembrá-las de que somos fortes e que podemos qualquer coisa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por sua bondade imerecida, pelo dom da vida, pela oportunidade de concluir essa inesquecível experiência acadêmica e por se tornar alguém tão real e íntimo na minha vida, por me acalmar, me dar forças e sabedoria nos momentos difíceis e de desânimo.

Agradeço a todos os professores e servidores do IFSP Campus Campinas, que contribuíram direta e indiretamente para a conclusão desse trabalho; em especial ao meu orientador Samuel, por sua inteligência e maestria, por me conduzir até aqui, por acreditar no meu potencial e me incentivar pessoal e profissionalmente em nossas reuniões diárias. Bem como ao professor Ubaldo, por sua parceria, disposição, interesse e visão detalhista.

Agradeço também à minha família, em seu sentido mais literal e amplo. A minha mãe Fran, meu exemplo, e meu padrasto Nilson, por todo o suporte desde o primeiro dia de faculdade, e por todo amor e apoio que demonstraram para que eu chegasse até aqui. É graças a eles. Também a minha avó, Vera, por seu carinho e compreensão.

Agradeço, ainda, ao meu amigo Lucas, por seu incentivo e preocupação em meio a correria dos semestres; aos meus parceiros de curso: Deive, Carol, Letícia, Ewerton, Giovanni, Marco, Aline e Vytor, pela honra do companheirismo dentro e fora dos corredores do campus; aos meus amigos do dia a dia: Mari, Gabriel, Rafa e João, pelos momentos de respiro, risadas, reflexões e episódios de Star Wars.

E para todas e todos que também participaram dessa trajetória, o meu muito obrigada!

"O ambiente escolar é um local que exerce influência intelectual e cidadã sobre um indivíduo, vindo a afetar a formação da identidade dos alunos. Identidade a qual é definida pelos comportamentos, atitudes e costumes de um indivíduo e se modifica com a convivência entre sujeitos, ou seja, se constrói tendo o Outro como referência."

Nilma Lino Gomes

RESUMO

O *bullying* é considerado um problema de saúde pública, sendo fator estimulante de diversos transtornos emocionais e comportamentais, assim como a diminuição do desempenho acadêmico. O presente trabalho tem o propósito de investigar técnicas de aprendizado de máquina para a predição de notas de alunos do ensino médio a partir de indicadores escolares e de *bullying*. Avaliamos quatro diferentes tipos de algoritmos regressores de aprendizado de máquina: *Linear Regression, Decision Tree, Random Forest* e *XGBoost*. O algoritmo *Random Forest* obteve os melhores resultados, com uma menor raiz quadrática média de erros, comparado ao segundo melhor algoritmo. Considerando as variáveis mais relevantes para o treinamento deste modelo, observamos que ter uma relação regular com os professores foi a que mais influenciou nas notas dos alunos.

Palavras-chave: *bullying*; desempenho escolar; aprendizado de máquina.

ABSTRACT

Bullying is considered a public health problem, being a stimulating factor of several emotional and behavioral disorders, as well as the decrease in academic performance. The present work aims to investigate machine learning techniques for predicting grades of high school students from school indicators and bullying. We evaluated four different types of machine learning regression algorithms: Linear Regression, Decision Tree, Random Forest, and XGBoost. The Random Forest algorithm obtained the best results, with a lower mean square root of errors, compared to the second best algorithm. Considering the most relevant variables for the training of this model, we observed that having a regular relationship with teachers was the one that most influenced students' grades.

Keywords: bullying; academic performance; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A partir do dia 7 de abril de 2016, o Brasil instituiu o Dia Nacional de Combate ao *Bullying* e à Violência na Escola com o objetivo de chamar a atenção para os reais problemas causados por este fato social, que atinge estudantes dentro e fora do espaço escolar. O *bullying* é considerado problema de saúde pública, visto que acarreta diversos transtornos emocionais e comportamentais, tais como: a diminuição da performance acadêmica (baixo desempenho escolar), abandono escolar, depressão, baixa auto-estima, ansiedade e, em casos mais severos, o suicídio (PIGOZI & MACHADO, 2015). Segundo Werf (2014 apud RIZZOTTO & FRANÇA, 2021, p. 2), o *bullying* não é um fenômeno isolado, pois a maioria das escolas no mundo vivencia esse tipo de violência. Ainda segundo a autora, a relação negativa entre o *bullying* e o desempenho acadêmico se dá pelo fato da vítima ter uma diminuição na frequência escolar, um menor contato com os colegas e perder motivação para estudar. O baixo desempenho escolar é um dos mais graves problemas que o sistema educacional enfrenta ao longo dos anos, percebido já nos anos iniciais da escolarização e que pode repercutir em prejuízos em diversas áreas, bem como para o próprio estudante vítima do *bullying* (SOUZA; ORELLANA; LEIVAS).

Uma pesquisa realizada com estudantes do ensino fundamental de escolas públicas de Recife em 2013 avaliou o efeito do *bullying* nas notas de matemática (OLIVEIRA et al., 2016). A metodologia empregada pelos autores foi o *Propensity Score Matching* (PSM) utilizando uma função de X, que estima o efeito médio do tratamento sobre os tratados, com o objetivo de comparar o efeito do *bullying* na performance de matemática dos alunos que alegaram ter sofrido *bullying* com um grupo de alunos que não sofreram *bullying*. Assim, de acordo com os autores, a função representa o *score*, ou probabilidade, que significa a probabilidade, neste caso, de sofrer *bullying*. Ainda segundo Oliveira et al., (2016), para calcular o efeito do *bullying* sobre a nota, foram usadas diversas técnicas estatísticas dos métodos de estimação. Os métodos de estimação são todas as informações obtidas através de dados amostrais, ou seja, que não abrangem o todo da população de interesse. Já que não é possível, por exemplo, calcular a média da altura de todos os seres humanos adultos existentes, é feito uma estimativa dessa média (HELENA, 2019). Os resultados obtidos por Oliveira et al., (2016), revelaram que o *bullying* tem um impacto negativo no desempenho dos alunos na disciplina, mais evidente entre as meninas. Além disso, os estudantes que sofreram

bullying apresentaram um desempenho inferior de aproximadamente 4,34% menor do que aqueles que não sofreram *bullying*. Da mesma forma, estudos de Rizzotto e França (2021) revelaram, também baseados pela metodologia *Propensity Score Matching* (PSM) e pelo Efeito Quantílico de Tratamento (EQT), que o *bullying* físico causa uma redução no desempenho escolar dos alunos, enquanto o *bullying* verbal não afetou negativamente a nota dos alunos. Ainda segundo Rizzotto e França (2021), os estudantes que apanharam tiveram suas notas reduzidas em 12, 33 e 20 pontos nas disciplinas de matemática, leitura e ciência respectivamente. Através do EQT, foi possível analisar os impactos do *bullying* na nota dos alunos.

Como observado, ambos os estudos obtiveram seus resultados por meio de métodos estatísticos e matemáticos e, graças ao aumento do poder de processamento dos computadores, as abordagens estatísticas e matemáticas evoluíram bastante, tornando possível combinar conhecimentos em estatística com avanços da computação (OLIVEIRA, 2020). Uma das áreas de grande interesse atualmente é a área de Ciência de Dados, que visa o estudo e análise de dados para a extração de conhecimento, sendo resultado da junção, principalmente, entre matemática e estatística e ciência da computação (DADOS; Ciência e., 2021). Uma de suas sub-áreas de maior interesse é o Aprendizado de Máquina (do inglês, *Machine Learning*), um ramo da inteligência artificial que visa o desenvolvimento de modelos computacionais que aprendem a reconhecer padrões e tomar decisões automaticamente a partir de dados. Um modelo de aprendizado de máquina aprende a partir de um conjunto de dados para resolver um problema sem ser explicitamente programado para isso. Dentre os tipos de problemas atacados com algoritmos de *machine learning*, vamos destacar o que será utilizado para a realização deste trabalho: regressão.

Um problema de **regressão** visa aprender um modelo que prediga um valor contínuo (real) ao invés de uma classe (categoria). Esse valor contínuo poderia ser, por exemplo, o preço de um imóvel ou de um produto, o peso ou altura de uma pessoa, etc. Há, ainda, técnicas de aprendizado de máquina que exploram **quais são as características** do problema que têm mais importância no resultado final obtido, isto é, as características de maior impacto no resultado da predição. Tais técnicas podem ser úteis no contexto do *bullying* escolar, uma vez que as mesmas poderiam apontar quais são os indicadores associados ao *bullying* que mais impactam as notas obtidas pelos alunos.

Até onde sabemos, há poucos trabalhos que utilizam aprendizado de máquina no contexto de *bullying*. Na verdade, de forma geral, os estudos sobre violência escolar são recentes e os primeiros tiveram início na década de 1980. De acordo com FANTE e PEDRA (2008 apud SILVA, 2015), o tema chegou ao Brasil no fim dos anos de 1990 e início de 2000. Tendo em vista que no Brasil os estudos sobre os impactos do *bullying* no desempenho escolar são recentes, a maioria dos brasileiros desconhece sua gravidade na vida dos estudantes. A falta de literatura em língua portuguesa com dados brasileiros acerca do tema também é uma causa (SILVA et al., 2019). Para Ponzo (2012), a literatura especializada é bastante ampla quando as discussões envolvem avaliar as causas do comportamento violento dos agressores e as consequências nos traços psicológicos das vítimas; porém, por outro lado, é escassa a quantidade de trabalho que têm abordado o efeito do *bullying* no desempenho acadêmico. De acordo com a Pesquisa Internacional sobre Ensino e Aprendizagem (IBGE, 2019), em 2019, o ambiente escolar brasileiro é duas vezes mais suscetível ao *bullying* do que a média geral das instituições de ensino em 48 países, apontando, também, a importância de medidas mais enérgicas contra o *bullying* no ambiente escolar. Ainda, de acordo com Batsche (1997 apud RIZZOTTO & FRANÇA, 2021, p. 2), o *bullying* é uma das maneiras mais recorrentes de violência nas escolas.

Para Neto (2005), o *bullying* é um problema universal encarado como natural, onde é frequentemente ignorado ou não valorizado pelos adultos. Dessa forma, nota-se a importância deste assunto para a sociedade e da urgência em investigar a influência de indicadores de *bullying* no desempenho escolar, principalmente por serem poucos os trabalhos que utilizam de aprendizado de máquina no contexto de *bullying*. Desta maneira, este trabalho objetiva a aplicação de técnicas de aprendizado de máquina no contexto da análise de indicadores de *bullying* no contexto escolar. Ou seja, será realizada a criação de modelos de aprendizado de máquina para prever as notas, entender as características mais importantes no resultado final obtido e avaliar os resultados.

2 OBJETIVOS

2.1 Objetivo geral

Este trabalho tem por objetivo a investigação e a criação de modelos de aprendizado de máquina para a predição de notas escolares de alunos do ensino médio e técnico integrado a partir de indicadores escolares, incluindo indicadores de *bullying*.

2.2 Objetivos específicos

a) Realizar um levantamento bibliográfico de trabalhos que investigam a influência de indicadores de *bullying* no desempenho escolar;
b) Buscar uma base de dados relacionado ao problema proposto;
c) Organizar, limpar e pré-processar a base de dados escolhida;
d) Investigar algoritmos com abordagens supervisionadas de *machine learning*;
e) Investigar algoritmos de regressão para a predição das notas dos alunos,
f) Analisar os atributos mais relevantes para a predição das notas feitas pelos tipos de regressores estudados.

3 ORGANIZAÇÃO DO TRABALHO

Nesta seção, apresentamos a organização da monografia, que é dividida como segue. O Capítulo 5 apresenta um levantamento bibliográfico de trabalhos relacionados que investigam a influência de indicadores de *bullying* no desempenho escolar. O Capítulo 6 apresenta os conceitos essenciais para a fundamentação teórica deste trabalho, enquanto o Capítulo 7 apresenta a base de dados utilizada para este trabalho. O Capítulo 8 apresenta a metodologia e os experimentos propostos, detalhando as etapas. Finalmente, o Capítulo 9 apresenta os resultados obtidos e o Capítulo 10 apresenta as conclusões deste trabalho, bem como possibilidades de trabalhos futuros.

4 TRABALHOS RELACIONADOS

Para Ponzo (2012), a literatura especializada é bastante ampla quando as discussões envolvem avaliar as causas do comportamento violento dos agressores e as consequências nos traços psicológicos das vítimas. Porém, por outro lado, é escassa a quantidade de trabalho que têm abordado o efeito do *bullying* no desempenho acadêmico. Ainda, segundo a autora que analisou o efeito do *bullying* no desempenho escolar de estudantes na Itália, aponta que ter sofrido *bullying* reduz as notas dos estudantes italianos.

Pereira et al. (2004 apud RIZZOTTO & FRANÇA, 2021, p. 4), analisaram, também, o *bullying* nas escolas portuguesas em alunos de 10 a 12 anos. Os resultados mostraram que o gênero tem um papel significativo nas vítimas, concluindo que ser do sexo masculino aumenta o risco de sofrer *bullying*.

De acordo com um estudo feito por Rigby e Slee (1991 apud RIZZOTTO & FRANÇA, 2021, p. 4) nas escolas australianas, crianças mais novas tendem a sofrer mais *bullying* do que as mais velhas, assim como os meninos foram intimidados mais do que as meninas.

Costa e Pereira (2010) analisaram a prevalência do *bullying* em 3.891 alunos com sucesso ou insucesso escolar, nos diferentes níveis do ensino básico em Portugal. Os alunos com insucesso, se comparado aos com sucesso escolar, se envolvem em mais episódios de *bullying*, onde ser vítima resulta em um desempenho escolar inferior.

Vignoles e Meschi (2010) apontaram que, no Reino Unido, os alunos que sofrem *bullying* têm níveis mais baixos de desempenho acadêmico e níveis mais baixos de prazer na escola. Já Kibriya, Xu e Zhang (2015 apud OLIVEIRA, et al., 2016, p. 5) analisaram o *bullying* escolar em 7.323 alunos de Gana. Os resultados mostraram o impacto negativo do *bullying* sobre a nota de matemática.

Observa-se, assim, uma regularidade na literatura internacional, que constata que o *bullying* impacta negativamente o desempenho escolar. Os estudos sobre *bullying* no Brasil são mais recentes. Segundo uma das pioneiras do assunto, Cléo Fante, o tema chegou ao Brasil no fim dos anos de 1990 e início de 2000, graças a uma grande pesquisa desenvolvida pela Noruega em 1980, expandindo os estudos para inúmeros países europeus (SILVA, 2015).

Malta et al. (2014, apud RIZZOTTO & FRANÇA, 2021, p. 5) avaliaram os indicadores associados ao bullying por meio da Pesquisa Nacional de Saúde do Escolar

(PeNSe), uma pesquisa em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE) e com o apoio do Ministério da Educação (MEC). Os autores encontraram que estudantes do sexo masculino de cor preta tem maiores chances de sofrer *bullying*.

Oliveira et al., (2016) realizaram uma pesquisa com estudantes do ensino fundamental de escolas públicas de Recife e avaliaram o efeito do *bullying* nas notas de matemática. Os resultados revelaram que o *bullying* tem um impacto negativo no desempenho dos alunos na disciplina, mais evidente entre as meninas. Além disso, os estudantes que sofreram *bullying* apresentaram um desempenho inferior de aproximadamente 4,34% menor do que aqueles que não sofreram *bullying*.

Silva et al. (2018 apud RIZZOTTO & FRANÇA, 2021, p. 5) encontraram uma relação negativa entre a escolaridade da mãe e a vitimização por *bullying*, de que a escolaridade avançada da mãe faz com que ela saiba impor limites, supervisione e auxilie os filhos quando estão com dificuldades na escola. Os autores ainda encontraram que as vítimas de *bullying* se sentem mais sozinhas, possuem menos amigos e têm dificuldade para dormir.

É por isso que a prevenção ao *bullying* deve ser tratada como um fenômeno sociocultural e que precisa ser discutida.

5 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresentará os conceitos relacionados ao *bullying*, bem como uma contextualização sobre aprendizado de máquina.

5.1 *BULLYING*

O primeiro pesquisador que percebeu o fenômeno *bullying* foi o professor de psicologia Dan Olweus em 1973. Seus estudos sobre o tema realizados na Universidade de Bergan (1978 a 1993), na Noruega, obtiveram uma grande repercussão. Porém, foi só depois do suícidio de três crianças entre 10 e 14 anos, vítimas de *bullying*, que o governo norueguês se atentou ao real problema e criou, assim, em escala nacional, a Campanha Anti-*Bullying* nas escolas em 1993 (QUINTANILHA, 2011). Essa Campanha Norueguesa Anti-*Bullying* reduziu índices de *bullying* e evasão escolar no país, viabilizando a melhora no desempenho acadêmico dos estudantes noruegueses.

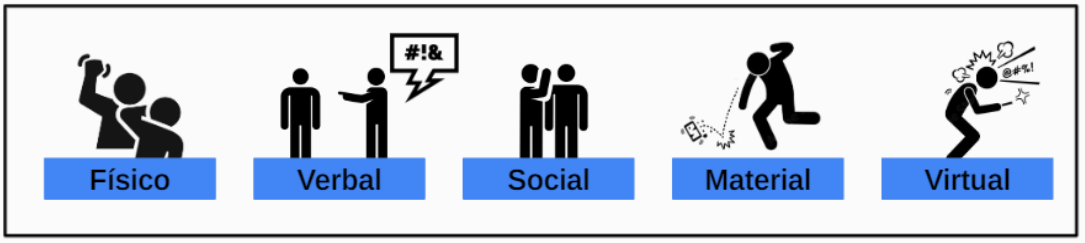
Para a autora e uma das pioneiras do assunto no Brasil, Cléo Fante, os estudos levantados e realizados pelo pesquisador Olweus são de grande relevância e importância, pois ele foi o responsável por desenvolver os primeiros critérios para detectar o problema de forma específica (2005 apud QUINTANILHA, 2011, p. 37).

O *bullying* (do inglês, *bully* = valentão, brigão) é um termo de origem inglesa, compreendido universalmente e popularizado pelo professor de psicologia Dan Olweus em 1973. O *bullying* é uma agressão física e psicológica, caracterizado por comportamentos agressivos intencionais e repetitivos, que acarretam diversos transtornos emocionais e comportamentais em quem sofre, como a diminuição da performance acadêmica (baixo desempenho escolar), abandono escolar, depressão, baixa auto-estima, ansiedade e, em casos mais severos, o suicídio (PIGOZI & MACHADO, 2015).

Ainda segundo o autor Alan Beane (2005 apud QUINTANILHA, 2011, p. 39), o termo descreve uma variedade de comportamentos que podem ter impacto sobre a propriedade, o corpo, os sentimentos, os relacionamentos, a reputação e o status social de uma pessoa.

A Figura 1 a seguir ilustra as diferentes formas que o *bullying* pode assumir.

Figura 1 - Tipos de *bullying*



Fonte: Elaborado pelo autor

Como pode ser observado na Figura 1, o *bullying* pode assumir cinco diferentes formas, sendo elas: *físico*, *verbal*, *social*, *material* e *virtual*.

O *bullying* **físico** é caracterizado por agressões físicas como socos, chutes e empurrões. O **verbal** por xingamentos, provocações depreciativas, apelidos ofensivos, intimidação e até ameaças. O **social** é quando há exclusão social intencional, por compartilhar mentiras, fofocas e rumores falsos. O **material** é o dano, a destruição ou o furto de pertences. O **virtual** - ou *cyberbullying*, ocorre na internet, fora do espaço escolar e dentro do virtual, por meio de redes sociais e sites (SILVA et al., 2014). Apesar de sua origem inglesa não ter uma tradução na língua portuguesa, é normal encontrar variações como intimidação, violência e agressividade.

Além disso, um estudo realizado com comportamentos de *bullying* entre jovens dos EUA (NANSEL et al., 2001) sugere que o efeito negativo do *bullying* é associado a um pior ajuste psicossocial, ou seja, aqueles que sofreram *bullying* demonstraram pior ajuste social e emocional, relatando dificuldade em fazer amigos e maior solidão.

De acordo com um estudo feito Rigby e Slee (1991 apud RIZZOTTO & FRANÇA, 2021, p. 4) nas escolas australianas, crianças mais novas tendem a sofrer mais *bullying* do que as mais velhas, assim como os meninos foram intimidados mais do que as meninas. Silva et al. (2018 apud RIZZOTTO & FRANÇA, 2021, p. 5) aponta que a escolaridade avançada da mãe faz com que ela saiba impor limites, supervisione e auxilie os filhos quando estão com dificuldades na escola. Os autores ainda encontraram que as vítimas de *bullying* se sentem mais sozinhas, possuem menos amigos e têm insônia.

Para Neto (2005), o *bullying* é problema não só do sistema educacional, mas de saúde pública, pois afeta as vítimas em níveis cognitivos e psicológicos, que podem perdurar na sua fase adulta e no trato com outros indivíduos. Ele é uma forma de afirmação de poder através da agressão.

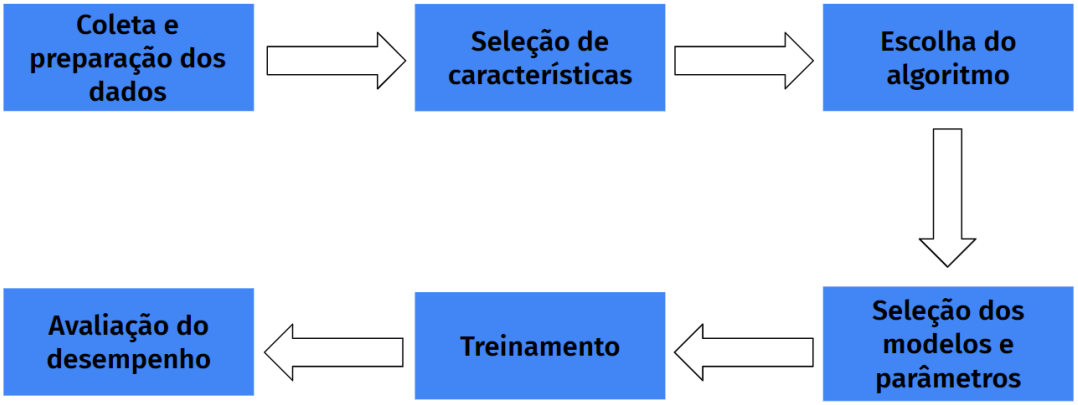
5.2 Aprendizado de máquina

O termo em inglês *machine learning* foi criado pelo engenheiro americano Arthur Samuel (HELDER, 2018). De acordo com o criador, [o aprendizado de máquina] "é o campo de estudo que dá ao computador a habilidade de aprender, sem ser explicitamente programado", pois nada mais é do que um método de análise de dados que, por meio de algoritmos, é capaz de encontrar e aprender padrões presentes nos dados recebidos.

Ainda segundo Brink e Richards (2014 apud CARVALHO, 2014, p.33), o aprendizado de máquina é "uma abordagem guiada a dados para a resolução de problemas, na qual padrões ocultos em um conjunto de dados são identificados e utilizados para ajudar na tomada de decisões". Assim, o aprendizado de máquina é um ramo da Inteligência Artificial, que permite que computadores tomem decisões com a ajuda de algoritmos (sequência de instruções) com o mínimo de intervenção humana, podendo aprender sozinhos com seus erros e fazer previsões sobre os dados analisados. É a partir dos algoritmos de aprendizado de máquina e da base de dados que são criados os modelos. Esse modelo, com a ajuda dos algoritmos, é capaz de se adaptar, modificar e melhorar o seu comportamento e suas respostas. (ESCOVEDO, 2020). O modelo aprende (treinando, construindo, formulando) a partir de um conjunto de dados. Depois de aprender, o modelo é aplicado, pois agora é capaz de fazer uma estimativa (teste, predição) de valores desconhecidos, novos e nunca antes visto (ESCOVEDO, 2020).

A Figura 2 mostra que um modelo genérico de aprendizado de máquina consiste em seis componentes independentes.

Figura 2 - Componentes de um modelo genérico de aprendizado de máquina



Fonte: Adaptado e traduzido de Alzubi; Nayyar; Kumar, 2018.

Como observado na Figura 2, cada componente do modelo tem uma tarefa específica a ser realizada. São elas:

- **Coleta e preparação dos dados:** consiste em coletar e preparar a base de dados (do inglês, *dataset*) em um formato adequado e que possa servir de entrada para o algoritmo. Geralmente esses dados não são estruturados, o que requer a imprescindível etapa de limpeza, normalmente dividida em duas fases: (1) examinar os dados (identificar valores ausentes, duplicados ou irrelevantes) e (2) limpar os dados (de fato remover ou substituir/preencher esses valores) para que estejam adequados para o problema.
- **Seleção de características:** os dados que foram obtidos na etapa anterior podem conter vários recursos, mas nem todos são relevantes para o processo de aprendizagem. Nesse caso, é selecionado as características mais importantes para o objetivo do problema (ALZUBI; NAYYAR; KUMAR, 2018).
- **Escolha do algoritmo:** nem todos os algoritmos de aprendizado de máquina são destinados a todos os problemas. Dependendo do problema, alguns são mais adequados do que outros. Assim, escolher o melhor algoritmo de aprendizado de máquina é fundamental para se obter os melhores resultados (ALZUBI; NAYYAR; KUMAR, 2018).

- **Seleção dos modelos e parâmetros:** consiste na técnica do modelo e na configuração e ajuste do seu hiperparâmetro.
- **Treinamento:** depois de selecionar o algoritmo apropriado, o modelo precisa ser treinado (aprende), usando uma parte do conjunto de dados.
- **Avaliação do desempenho:** depois de treinar com dados rotulados, é hora do modelo pôr em prática o que aprendeu, contra dados nunca vistos antes. Essa avaliação é realizada por meio de métricas, que vão dizer se o desempenho foi bom ou ruim.

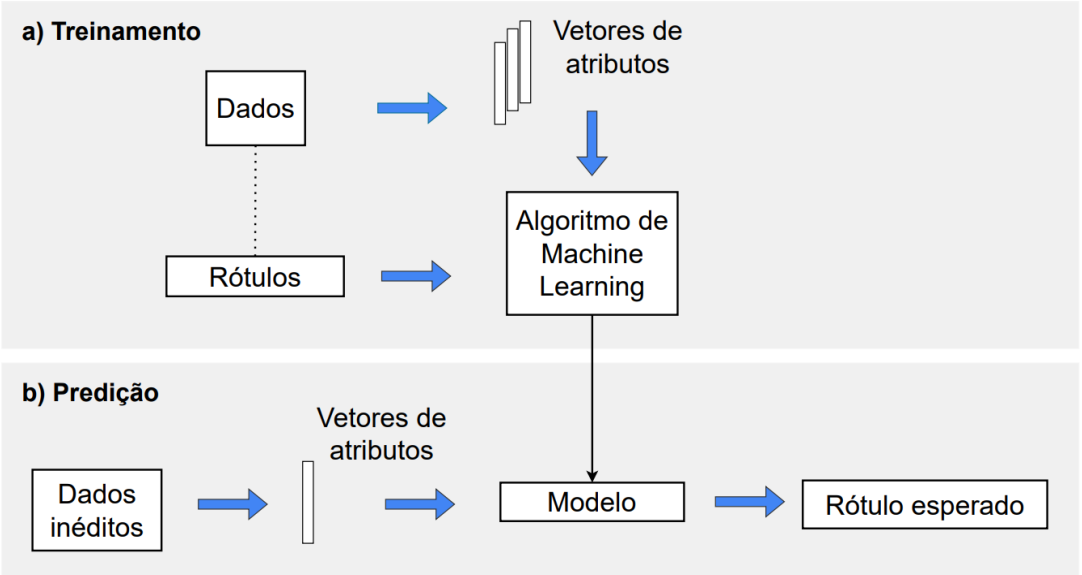
Dentre as muitas abordagens de aprendizado de máquina, destacam-se o aprendizado supervisionado e não supervisionado. A abordagem supervisionada (considerada em nosso trabalho na seção 5.2.1) consiste em o modelo executar uma tarefa a partir de uma série de exemplos (ou instâncias) previamente rotulados, ou seja, a partir das respostas corretas. Já na abordagem não supervisionada, o treinamento acontece em localizar padrões em dados que não são rotulados.

5.2.1 Aprendizado de máquina supervisionado

Dentre os diferentes tipos existentes de sistemas de aprendizado de máquina, a abordagem do aprendizado supervisionado consiste em ensinar para o computador a executar uma tarefa a partir de uma série de exemplos (ou instâncias) previamente rotulados, ou seja, a partir das respostas corretas (SILVEIRA & BULLOCK, 2017), daquilo que sabemos a saída esperada e que são passadas para o modelo anteriormente. Dessa forma, a partir da experiência obtida através desses exemplos, o computador aprende, reconhece um padrão e se torna capaz de fazer predições a partir de instâncias novas de maneira precisa, nunca vistos antes. Seu objetivo é predizer um valor.

Esse processo é ilustrado na Figura 3 a seguir.

Figura 3 - Funcionamento do aprendizado de máquina supervisionado



Fonte: Adaptado e traduzido de Alzubi; Nayyar; Kumar, 2018.

Na Figura 3-(a), o algoritmo de *machine learning* está sendo treinado por meio da entrada de dados de interesse. As entradas possuem rótulos que permitem a identificação desses dados dentro do conjunto de treinamento, que servirá para construir o modelo. Como observado, os dados e seus rótulos atuam juntos (formam os vetores de atributos) e são utilizados pelo algoritmo de *machine learning* para generalizar um modelo e permitir que este preveja o rótulo a partir de dados ainda não vistos (dados inéditos), a serem testados.

Já a etapa apresentada na Figura 2-(b), ilustra esta previsão, de modo que o algoritmo entenda os dados sendo testados e possa atribuir o rótulo esperado. Isto é, o algoritmo colocará em prática o que treinou, sendo capaz de prever as notas escolares de alunos a partir de indicadores escolares e de *bullying*. Entre as técnicas de aprendizado de máquina supervisionado, a **regressão** é uma tarefa típica, que consiste em prever ou estimar um valor numérico. Geralmente, as técnicas de regressão consistem em encontrar uma relação entre um conjunto de dados de entrada e uma saída numérica (GÉRON, 2019). Se as instâncias forem valores contínuos, temos um problema de regressão, como, por exemplo, prever o preço de um carro a partir de um conjunto de características (quilometragem, marca, etc.)

Este trabalho tratará de um problema de regressão, visto que o objetivo é a predição de notas de alunos a partir de indicadores escolares e de *bullying*.

5.2.2 Treinamento, teste e validação

Para a resolução dos mais diversos problemas utilizando algoritmos de aprendizado de máquina, é importante que a máquina aprenda e consiga aplicar o que aprendeu. Por isso, separamos a totalidade da base de dados em dois grupos, sendo o primeiro responsável pelo aprendizado (treinamento) do modelo e o segundo para realizar os testes.

Os dados de treino são os dados que serão apresentados ao algoritmo de aprendizado de máquina para a criação dos modelos, previamente rotulados (DIDÁTICA, 2019). Enquanto os de teste serão apresentados após a criação do modelo, usado para validar o desempenho real do modelo construído, visto que os resultados obtidos serão comparados com os valores reais (DIDÁTICA, 2019).

Segundo ESCOVEDO (2021), para treinar e validar o modelo de aprendizado de máquina supervisionado, uma das alternativas é utilizar a estratégia *train-test-split*, que justamente consiste em dividir a base de dados em dois subconjuntos menores para serem utilizados como dados de treino e teste (ou validação). Essa função nos garante que os dados serão aleatoriamente distribuídos entre os dois grupos, onde é possível informar o tamanho de cada grupo.

5.3 Pré-processamento de dados

Na literatura especializada, a etapa de pré-processamento é um conjunto de atividades que envolvem a preparação, organização e estruturação de um conjunto de dados. Sua importância se dá para garantir a qualidade e confiabilidade das conclusões (PINHEIRO, 2021. Esta seção aborda as técnicas de pré-processamento para o desenvolvimento deste trabalho.

5.3.1 Tipos de dados

Um conjunto de dados pode conter diferentes tipos de variáveis. Existem duas grandes classes de variáveis, sendo elas: variáveis quantitativas e qualitativas.

As variáveis quantitativas referem-se exclusivamente a valores numéricos, que se subdividem em contínuas e discretas. As contínuas resultam de um número infinito de valores possíveis que podem ser associados em uma escala contínua, como, por exemplo, o peso de um pacote de arroz. As discretas resultam de um conjunto finito de valores possíveis, geralmente são o resultado de contagens, como a quantidade de livros de Aprendizado de Máquina da biblioteca da escola (SILVIA, 2012).

Já as variáveis qualitativas (ou categóricas) indicam a qualidade de algo, que representa classificação de indivíduos, que se subdivide em nominais e ordinais. As nominais é quando não existe ordenação entre as categorias, como, por exemplo, sexo ou a cor dos olhos. As ordinais é quando existe ordenação entre as categorias, como escolaridade ou um *ranking* de um jogo (SILVIA, 2012).

Por entender o tipo de variável presente na base de dados, fica fácil saber o que fazer com cada informação e qual o melhor tipo de tratamento a ser aplicado.

5.3.2 Limpeza de dados

O conjunto de dados original pode conter dados irrelevantes ou ausentes, visto que geralmente esses dados não são estruturados. Para lidar com essa situação, a limpeza de dados é essencial, normalmente é dividida em duas fases: (1) examinar os dados (identificar valores ausentes, duplicados ou irrelevantes) e (2) limpar os dados (de fato remover ou substituir/preencher esses valores) para que estejam adequados para o problema.

Para dados faltantes, ou seja, quando alguns dados estão ausentes, existem técnicas para se resolver, como remover registros com atributos nulos ou realizar uma média ou mediana com os valores do mesmo atributo (CÉSAR, 2019).

5.3.3 Técnicas de transformação

Levando em consideração que as variáveis vistas na seção 5.3.1 são utilizadas como valores de entrada pelos algoritmos de aprendizado de máquina, é necessário, geralmente, representar numericamente qualquer variável que não seja numérica, uma vez que a maioria dos modelos de aprendizado de máquina exige que todas as variáveis estejam em formato numérico (MIRANDA, 2022).

Há várias técnicas de transformação e codificação para lidar com variáveis categóricas, visto que não podem ser utilizadas em forma de texto no momento de treinar os modelos de aprendizado de máquina.

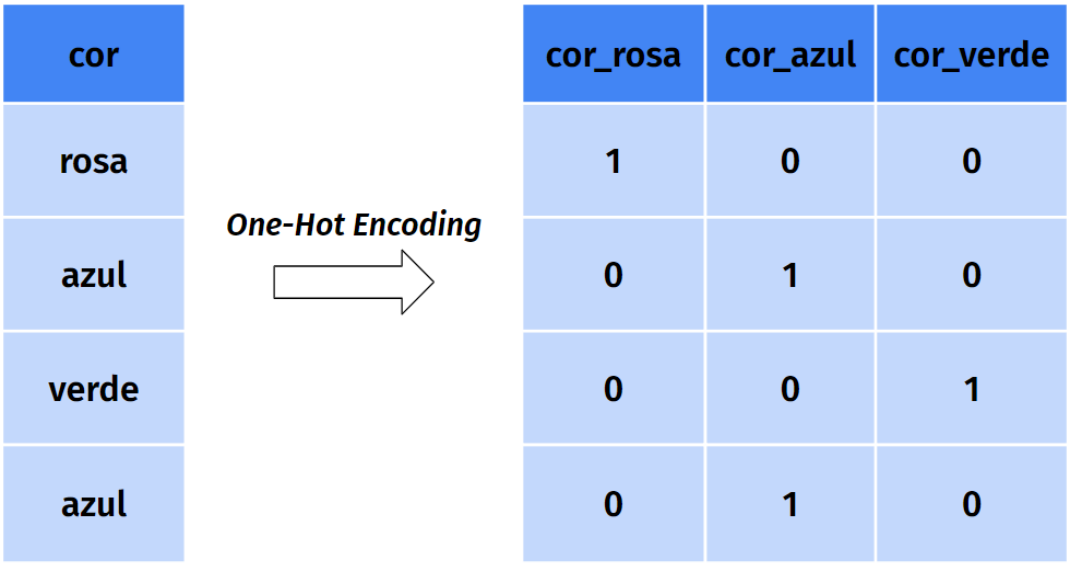
A seguir, será abordado duas técnicas para o desenvolvimento deste trabalho.

5.3.3.1 *One-hot encoding*

O processo de transformação das variáveis categóricas usando o *One-Hot Encoding* é feito a partir da criação de novas colunas a partir das categorias. Essa técnica cria novas colunas binárias para cada categoria, onde, caso tenha a presença da característica, o valor correspondente será 1. Caso contrário, 0, sendo a ausência da característica.

A Figura 4 a seguir ilustra como esse processo funciona.

Figura 4 - Funcionamento da técnica *One-Hot Encoding*



Fonte: Adaptado e traduzido de Novack, 2020.

Ou seja, é criado N novos recursos, onde N é o número de valores exclusivos no conjunto original (primeira tabela). Na Figura 4, o N seria igual a três, pois existem três cores únicas (rosa, azul e verde). Após a aplicação da técnica, é possível observar que o primeiro registro, por exemplo, é da cor rosa, pois contém o valor 1 (indica presença) na coluna ***cor_rosa***, e 0 (indica ausência) nas colunas ***cor_azul*** e ***cor_verde***.

5.3.3.2 *Standard scaler*

O *Standard Scaler* trabalha com valores numéricos, os colocando na mesma escala. Para evitar erros nas previsões, essa técnica padroniza as características, o que os torna mais manejáveis para os modelos de aprendizado de máquina. Ou seja, para cada característica, a média é 0 e o desvio padrão 1 (RAFAEL, 2020). A padronização de um conjunto de dados é um requisito comum para algoritmos de aprendizado de máquina, visto que se não estiverem no mesmo padrão, podem apresentar comportamentos indesejáveis. (RAFAEL, 2020).

5.4 Algoritmos de aprendizado de máquina

Esta seção aborda os conceitos por trás dos algoritmos selecionados para o desenvolvimento deste trabalho, sendo todos algoritmos de regressão.

5.4.1 *Linear Regression*

A regressão linear (do inglês, *linear regression*) é um dos algoritmos mais conhecidos em estatística e aprendizado de máquina, que busca estabelecer a relação linear entre uma variável dependente e uma ou mais as variáveis independentes. A variável independente (ou preditora) é aquela que será passada previamente para o modelo. Enquanto a variável dependente (ou alvo) é aquela que queremos prever (DAMACENO, 2020).

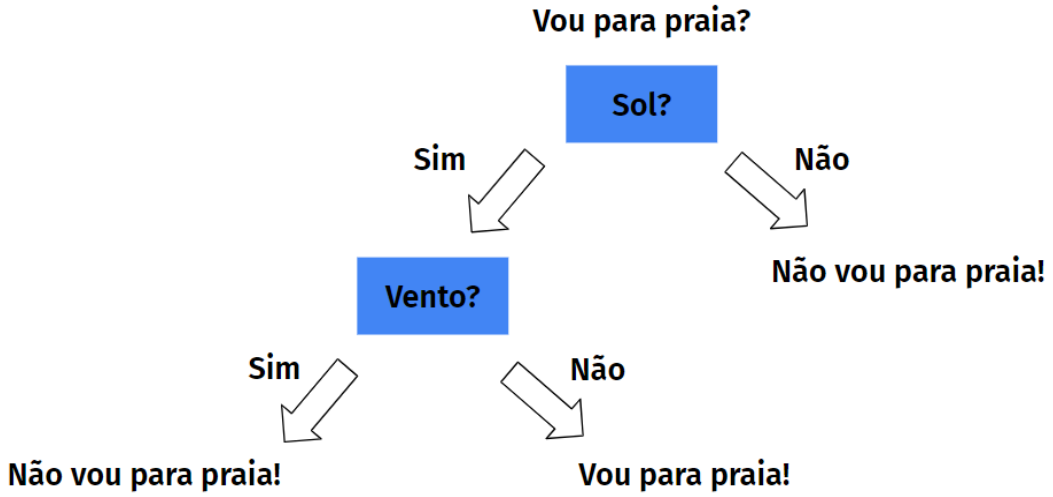
Essa relação é representada por um modelo matemático que estima os coeficientes da equação linear que melhor preveem o valor da variável dependente (SOUSA, 2022). Ainda de acordo com Sousa (2022), tem-se na regressão um fenômeno de interesse e um número de

observações, onde cada um possui uma ou mais características. Assumindo que, pelo menos, uma das características depende de outras, tenta-se encontrar um relacionamento entre elas.

5.4.2 *Decision Tree*

A árvore de decisão (do inglês, *decision tree*), conforme o nome sugere, vários pontos de decisão são criados. Esses pontos são os "nós" da árvore e em cada um deles o resultado da decisão será seguir por um caminho, ou outro. Esses caminhos existentes são os ramos. É, basicamente, uma representação de um conjunto de regras criado para tomar qualquer decisão, nesse caso, estimar um valor; ou seja, uma pergunta será feita e haverá duas opções de resposta: sim ou não, onde uma opção "sim" leva a próxima pergunta, e a opção "não" a outra (DIDÁTICA, 2020). A Figura 5 a seguir ilustra a estrutura de uma árvore de decisão.

Figura 5 - Estrutura de uma árvore de decisão



Fonte: Adaptado e traduzido de Didática, 2020.

O exemplo da Figura 5 acima é bem rudimentar, mas é facilmente compreendido. Como pode ser observado, o algoritmo criará uma estrutura parecida a um fluxograma com

"nós", onde uma condição é verificada, e se atendida, o fluxo segue por um ramo, caso contrário, por outro, sempre levando ao próximo nó, até a finalização da árvore.

Ainda de acordo com DIDÁTICA (2020), para os problemas de regressão, o objetivo é prever um valor. Para isso, a árvore utilizará os conceitos de média e desvio padrão, que possibilita um resultado final numérico. Para definir as variáveis independentes dos nós principais em um problema de regressão, será calculado o desvio padrão dos valores da variável dependente para cada variável independente, de acordo com suas variações. Assim, a variável independente com maior redução de desvio padrão (indica o quão distante os valores estão da média) será escolhida para o nó principal da árvore.

Com os dados de treino, o algoritmo busca as melhores condições, e onde inserir cada uma dentro do fluxo. Dessa forma, o algoritmo aprenderá os melhores critérios para dividir um ponto em dois nós, e decidirá onde cada ponto de dados deve ser incluído na árvore até chegar a uma decisão final.

5.4.3 *Random Forest*

A árvore de decisão vista na seção anterior é a base do funcionamento de outros algoritmos, como a árvore aleatória (do inglês, *random forest*). A árvore aleatória irá criar muitas árvores de decisão, de maneira aleatória, formando uma floresta, onde cada árvore será utilizada na escolha do resultado final, em uma espécie de votação (DIDÁTICA, 2020). Diferentemente do que acontece na criação de uma árvore de decisão simples, ao utilizar o RandomForest, o primeiro passo executado pelo algoritmo será selecionar aleatoriamente algumas amostras dos dados de treino, e não a sua totalidade (DIDÁTICA, 2020).

Os dados dessas árvores são então mesclados para garantir as previsões mais precisas. Enquanto uma árvore de decisão individual tem um resultado e uma gama limitada de grupos, a floresta garante um resultado mais preciso, com um número maior de grupos e decisões.

5.4.4 *XGBoost (eXtreme Gradient Boosting)*

O *XGBoost* é um dos algoritmos mais utilizados, baseado em árvore de decisão, apresentando resultados superiores, principalmente em problemas de previsão, que implementa um processo chamado *Boosting* para gerar modelos precisos. Nos últimos anos, o

XGBoost tem crescido em número de implementações, assumindo o posto de modelo de predição mais utilizado entre os vendedores de competições de aprendizado de máquina (CÉSAR, 2019).

Os algoritmos de *Boosting* combinam muitos modelos, utilizando um método mais inteligente, onde os novos modelos adicionados tem o objetivo de corrigir os erros dos modelos anteriores. Ou seja, O *XGBoost* busca construir a melhor árvore que vai minimizar os erros dos modelos anteriores (SCUDILIO, 2020).

5.5 Avaliação de desempenho do modelo

A avaliação do modelo é realizada por meio de métricas, que vão dizer se o desempenho foi bom ou ruim. Esta seção aborda a métrica utilizada para o desenvolvimento deste trabalho.

5.5.1 RMSE

A Raiz do Erro Quadrático Médio (do inglês, *Root Mean Squared Error - RMSE*) é uma métrica de avaliação muito utilizada e reconhecida para medir o desempenho de modelos de regressão. É a medida que calcula "a raiz quadrática média" dos erros entre valores observados (reais) e predições (hipóteses). O RMSE é calculado utilizando a seguinte equação.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

O RMSE é a diferença entre o valor que foi previsto pelo modelo e o valor real (rotulado, aquele que conhecemos) que foi observado, assim, é possível medir e analisar os

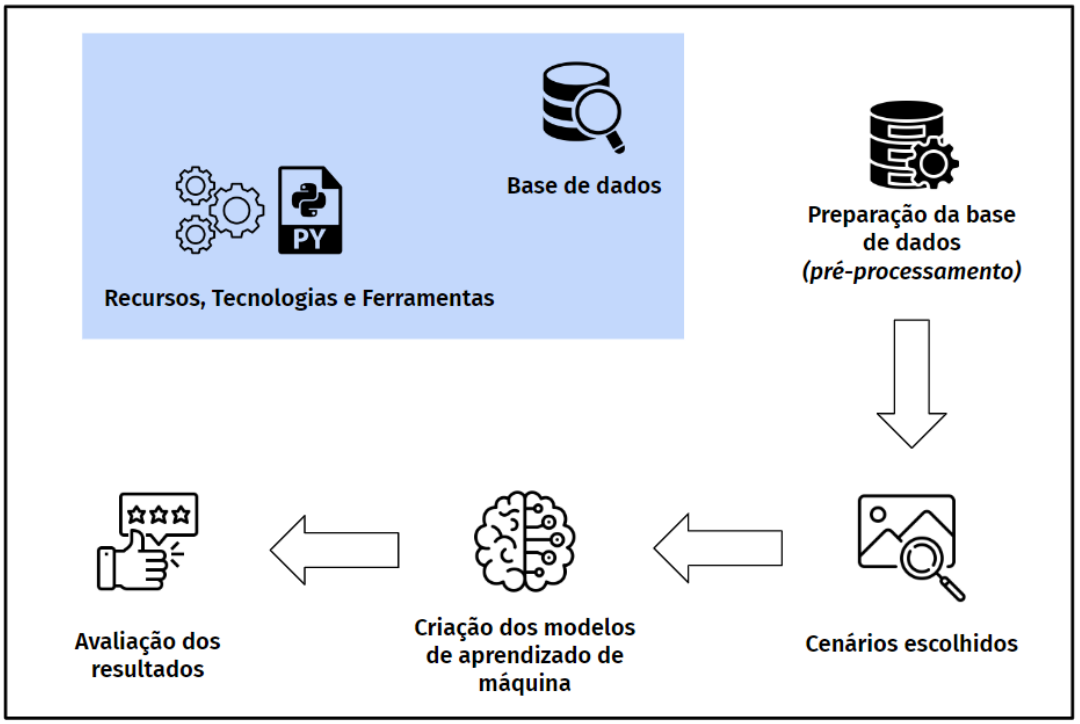
erros que ele apresenta, comparando os dados rotulados com os previstos (VASCONCELLOS, 2018). O **n** observado acima corresponde ao número de amostras, **yi** o valor real/observado da variável dependente para a amostra i e o **yi-hay (ŷ)** o valor previsto pelo modelo para a amostra i.

Quanto maior o RMSE do modelo, menor é o seu desempenho, já que valores altos indicam maior quantidade de erros de predição. O RMSE ajuda a saber quão bem o modelo é capaz de prever a resposta, e forma o critério mais importante para o ajuste quando o principal objetivo é usar o modelo para fins de previsão. Assim, essa medida penaliza predições muito distantes da real, ou seja, erros extremos (COSTA, 2019).

6 METODOLOGIA E EXPERIMENTOS

Os experimentos propostos para o desenvolvimento deste trabalho são ilustrados pela Figura 6 a seguir. As etapas são descritas nas subseções presentes nesta seção 8.

Figura 6 - Etapas da metodologia e experimentos



Fonte: Elaborado pelo autor

6.1 Recursos, tecnologias e ferramentas

Para o desenvolvimento deste trabalho, será necessário programas para manipulação, tratamento de dados e criação dos modelos. Consideraremos as seguintes tecnologias:

- Linguagem de programação ***Python***[1], devido ao seu alto poder de processamento, suporte de biblioteca, capacidade de análise estatística e simples codificação. Além de ser, atualmente, a linguagem mais recomendada para Ciência de Dados;
- Plataforma para codificação e desenvolvimento, ***Jupyter Notebook***[2];

[1] https://www.python.org/
[2] https://jupyter.org

- Bibliotecas ***Pandas***[3] e ***Numpy***[4], do *Python* para a manipulação e tratamento dos dados, comumente utilizadas na etapa de limpeza de dados, visto na seção 5.3.2;
- Biblioteca ***Scikit-learn***[5], desenvolvida especificamente para aplicação prática de aprendizado de máquina. Suas aplicações utilizadas neste trabalho são: pré-processamento, desenvolvimento de modelos por algoritmos de regressão (regressão linear, árvores de decisão e árvores aleatórias) e criação efetiva dos modelos de aprendizado de máquina.
- Biblioteca ***XGBoost***[6] , do algoritmo *XGBoost*.

6.2 Base de dados

Existem três tipos de dados: os *qualitativos* (indicam a qualidade de algo, como o tamanho ou cor); *quantitativos* (refere-se exclusivamente a números, cabendo perguntas como preço ou quantidade); e, por fim, os *categóricos* (identificados por categorias, como carros novos e usados). Para compor a base (ou conjunto) de dados relacionada ao problema proposto, será utilizado o estudo feito por Viera-Junior, Vieira, Moretti (2020), coletado entre agosto e setembro de 2019. Trata-se de uma pesquisa transversal realizada com estudantes adolescentes com idade entre 14 e 19 anos matriculados em escolas públicas e privadas do 1° ao 3° ano do ensino médio normal e técnico integrado na cidade de Campinas/SP, nos períodos matutino e vespertino.

Somando 489 amostras, esta pesquisa foi realizada por meio de um questionário onde os alunos presentes em sala de aula responderam a próprio punho, contendo perguntas quantitativas e qualitativas e estruturadas da seguinte forma:

- Relacionadas com caráter geral, pessoal e escolar, como:
    - Idade, gênero, cor, ano escolar, qualidade do sono, relação com amigos, relação com professores, classificação da auto-imagem corporal (como o aluno se vê em relação ao seu corpo). Com padrão de respostas de “muito ruim”, “ruim”, “regular”, “boa” e “muito boa”.
- Relacionadas a vitimização por *bullying*:

[3] https://pandas.pydata.org/
[4] https://numpy.org/
[5] https://scikit-learn.org/stable/
[6] https://xgboost.readthedocs.io/en/stable/

- Com 12 perguntas referentes a vitimização por *bullying* e *cyberbullying* com respostas de auto relato, como “nunca”, “uma vez” e “mais de uma vez”, durante o primeiro período letivo de 2019. As formas de *bullying* medidas foram: físico, social, verbal, material e virtual, abordadas na seção 5.1.

Paralelamente, foram coletadas, ainda, as notas oficiais dos alunos que participaram da pesquisa. Todas as respostas foram armazenadas em um arquivo Excel, contendo 24 colunas/atributos e 489 linhas/registros. A Tabela 1 a seguir simplifica o conteúdo da base de dados, contendo a variável e sua descrição.

Tabela 1 - Exemplo da base de dados no formato de planilha

Variável	Descrição
Escola	Unidade escolar.
Curso	Ensino médio normal ou técnico integrado.
Cor	Raça/cor do aluno.
Genero	Gênero do aluno.
Ano	Ano escolar do aluno.
Idade	Idade do aluno.
Demora dormir	Quantas vezes por semana demora a dormir.
Sonhos ruins	Quantidade de sonhos ruins.
Qualidade do sono	Classificação da qualidade do sono.
Relação com Amigos	Classificação da relação com os amigos.
Relação com Professores	Classificação da relação com os professores.
Auto-Imag	Classificação da auto-imagem.
Fisico	Quantidade de vitimização por bullying físico.
Social	Quantidade de vitimização por bullying social.
Verbal	Quantidade de vitimização por bullying verbal.
Material	Quantidade de vitimização por bullying material.
Virtual	Quantidade de vitimização por bullying virtual.
Mora com pais	Se mora ou não com os pais.
Escolari-mãe	Escolaridade da mãe.
Escolari-pai	Escolaridade do pai.
Nota-exatas	Nota na área de exatas.
Nota-humanas	Nota na área de humanas.
Nota-biológicas	Nota na área de biológicas.
Tipo escola	Escola privada ou pública.

Fonte: Elaborado pelo autor

Como exemplificado na Tabela 1, a base de dados é composta por variáveis que contém valores numéricos e, em sua maioria, categóricos. A coluna “Nota-exatas”, por exemplo, apresenta a nota do estudante na área de exatas e é composta por valores numéricos contínuos de 1 a 5, enquanto a coluna “Auto-Imag” é populada com valores categóricos ordinais, que indicam como o aluno se vê em relação ao seu corpo entre “muito ruim”, “ruim”, “regular”, “boa” e “muito boa”.

6.3 Preparação da base de dados

Nesta etapa, foi realizado um pré-processamento dos dados, que transforma dados brutos em formatos que sejam úteis e eficientes para uso e que serão utilizados na criação dos modelos preditivos. O carregamento desses dados, no formato de planilha *Excel*, será o primeiro passo para esta etapa, utilizando as bibliotecas *pandas* e *numpy* do *Python* para auxiliar na manipulação e tratamento desses dados.

Como a base de dados original continha algumas colunas redundantes e metodologicamente não muito interessantes, foi realizada a limpeza para adequação. Depois de examinar e identificar esses valores, foi feita a remoção destas colunas, bem como a criação de colunas novas com novos valores, geradas a partir das já existentes, até estarem adequadas para a necessidade e foco do problema.

Tabela 2 - Colunas originais com seus valores originais, colunas finais com seus valores finais

Variável Original	Valores Originais	Variável Final	Valores Finais
Cor	['Branca', 'Parda', 'Preta']	Cor Agrupada	['Branca', 'Não-Branca']
Nota-exatas	Nota na área de exatas (de 1 até 5).	Média-notas	[3, 3.33333333, 4, 3.66666667, 2.33333333, 5, 4.66666667, 2.66666667, 4.33333333, 1.66666667, 2]
Nota-humanas	Nota na área de humanas (de 1 até 5).		
Nota-biológicas	Nota na área de biológicas (de 1 até 5).		
Fisico	Quantidade de vitimização por bullying físico.	Bullying Total	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 25, 26]
Social	Quantidade de vitimização por bullying social.		
Verbal	Quantidade de vitimização por bullying verbal.		
Material	Quantidade de vitimização por bullying material.		
Virtual	Quantidade de vitimização por bullying virtual.		
		Sofreu Bullying?	['Não Sofreu Bullying', 'Sofreu Bullying']

Fonte: Elaborado pelo autor

Na Tabela 2, temos as colunas que passaram pelo processo de limpeza e seus valores antes e depois da mesma. Então, partindo do ponto de vista no que diz a literatura especializada, a partir das colunas já existentes, foram criadas novas colunas contendo:

- **Cor Agrupada**: raça/cor do aluno agora agrupadas entre branca e não branca (parda + preta);
- **Média-notas**: realizada a média das notas das três áreas do conhecimento.
- ***Bullying* Total**: soma de todas as formas de *bullying* para ter mais noção do impacto;
- **Sofreu Bullying?**: a partir da coluna *Bullying* Total, foi realizada criação da variável binária se sofreu ou não sofreu *bullying*; ou seja, para caso afirmativo aqueles que sofreram uma quantidade maior ou igual a um, e para caso negativo igual a zero.

Importante lembrar, ainda, que as colunas já existentes também serão utilizadas para a criação dos cenários comparativos. Depois da limpeza, partimos para o tratamento das variáveis, vista na seção 5.3.3, onde foram aplicadas duas técnicas de transformação: *One-Hot Encoding* para todas as variáveis categóricas da base de dados *e Standard Scaler* para todos os valores numéricos da base de dados.

6.4 Cenários escolhidos

Para cada algoritmo utilizado neste trabalho e a fim de explorar a correlação entre as variáveis e suas associações, foram criados dois cenários a partir dos indicadores de *bullying* presentes no conjunto de dados. Os dois cenários são: **(1)** se o aluno sofreu ou não *bullying* (variável binária); e **(2)** considerando toda a vitimização por *bullying* (todas as cinco formas abordadas na seção 5.1). Dentro de cada cenário, foram criados quatro grupos que trabalham com o mesmo tipo de variável, pois serão observados em conjunto.

O primeiro grupo contém todas as características do conjunto de dados. O segundo apenas as variáveis que dizem respeito ao caráter pessoal do aluno (cor/raça, gênero, auto-imagem e o tipo de escola). O terceiro grupo contém as relações (relação com os amigos e relação com os professores). O quarto, e último, diz respeito ao sono do aluno (demora para

dormir, quantidade de sonhos ruins, e a qualidade do sono). Esses grupos formam as variáveis independentes (preditoras), onde a dependente (alvo) é a média das notas.

Dessa forma, ficará mais tangível de explorar as variáveis e entender se existe alguma forma de *bullying* que mais se destaca ou se o simples fato do aluno já ser uma vítima deste fato já basta.

6.5 Criação dos modelos de aprendizado de máquina

A escolha dos algoritmos de aprendizado de máquina foi baseada no objetivo deste trabalho, que consiste em um problema de regressão, que é prever as notas escolares de alunos a partir de indicadores escolares e de *bullying* e entender as características mais importantes no resultado final. Os algoritmos utilizados são da abordagem supervisionada e do tipo regressão, sendo: *Linear Regression, Decision Tree, Random Forest* e *XGBoost*.

Referente à separação dos dados pré-processados nos dois conjuntos necessários para treinamento e teste, foi utilizado a estratégia *train-test-split*, apresentada na seção 5.2.2. Originalmente, a base de dados é composta por 489 linhas. A subdivisão desta base para compor os conjuntos de treinamento e teste, ficou da seguinte forma: os dados de treinamento receberão 80% dos dados (equivalente a 391 linhas), e os dados de teste 20%, que são 98 linhas.

6.6 Avaliação dos resultados

Para a análise e avaliação dos resultados das predições, foi utilizada a métrica explicada na seção 5.5.1, a Raiz do Erro Quadrático Médio (RMSE). Além disso, para realizar a interpretação do modelo, foi utilizado o ELI5[7], um pacote do *Python* que auxilia na visualização de quais são as variáveis mais importantes/relevantes para a predição. Ao entender as variáveis mais importantes, será possível selecionar os dados mais apropriados para os modelos, visto que, quando não apropriados, podem atrapalhar os resultados finais do aprendizado de máquina. Em um formato de tabela, o ELI5, através da função *show_weights*, calcula uma pontuação para todos os recursos de entrada para o modelo — as pontuações (ou

[7] https://eli5.readthedocs.io/en/latest/overview.html

os pesos) representam a importância de cada característica. Quanto maior o peso atribuído, mais importante é para o modelo; assim o ELI5 classifica as características com base no efeito que elas têm na previsão do modelo.

7 RESULTADOS

Neste capítulo, apresentaremos os resultados obtidos seguindo a metodologia apresentada no capítulo anterior. Os modelos preditivos foram criados com o objetivo de realizar a predição que informa as notas escolares dos alunos vítimas e não vítimas de bullying. Para o teste e validação do desempenho dos modelos, foi utilizado o método RMSE.

7.1 Resultados e análise do cenário 1

Nesta seção, apresentamos os resultados do cenário 1, contendo a variável binária se sofreu ou não sofreu *bullying*. A Tabela 3 a seguir apresenta os algoritmos utilizados e a classificação da métrica RMSE. Cada linha descreve um conjunto de variáveis considerado no treinamento dos modelos. Cada conjunto também possui a variável que define se o aluno sofreu ou não *bullying*.

Tabela 3 - Resultados das predições utilizando RMSE do cenário 1 para o conjunto de teste.

Sofreu ou não *bullying*

Variável	***Linear Regression***	***Decision Tree***	***Random Forest***	***XGBoost***
	RMSE	**RMSE**	**RMSE**	**RMSE**
Todas as variáveis	0.69	0.80	**0.58***	0.68
Caráter pessoal (cor/raça, gênero, auto-imagem e tipo de escola	0.61	0.68	**0.60***	0.68
Relação com amigos e professores	0.58	0.58	**0.57***	0.58
Sono (demora para dormir, quantidade de sonhos ruins e qualidade do sono	**0.58***	0.66	0.60	0.66

*Melhor resultado (menor RMSE) entre os modelos para um mesmo conjunto de variáveis.

Conforme a Tabela 3, podemos observar que ao considerar todas as variáveis e os agrupamentos/conjuntos de variáveis de caráter pessoal e de relações, o modelo que apresentou o menor erro (RMSE) para o teste foi o *Random Forest*, com um RMSE de 0.58, 0.60 e 0.57, respectivamente. Seu algoritmo possui diversos parâmetros para a construção da árvore. O *n_estimators*, por exemplo, define a quantidade de estimadores de árvores que serão usados para compor a "floresta", utilizamos 200 árvores, classificado como o melhor parâmetro. Por padrão, o algoritmo também utiliza *bootstraps*, amostras diferentes da base de dados, para construir as árvores, o que auxilia no aprendizado de hipóteses diferentes. Dentre os três melhores cenários, o conjunto de variáveis com relação com amigos e professores foi o destaque deste cenário, com RMSE de 0.57. Já no conjunto de variáveis de sono, o melhor modelo foi o *Linear Regression*, com RMSE de 0.58.

No geral, é possível perceber que o modelo *Random Forest* foi o que mais se destacou, dentre os demais, com o menor RMSE deste cenário.

7.2 Resultados e análise do cenário 2

Nesta seção, apresentamos os resultados do cenário 2, agora contendo todas as vitimizações de *bullying* (físico, verbal, social, material e virtual). A Tabela 4 a seguir apresenta os algoritmos utilizados e a classificação da métrica RMSE. Cada linha descreve um conjunto de variáveis considerado no treinamento dos modelos. Cada conjunto também possui a variável que define se o aluno sofreu ou não *bullying*.

Tabela 4 - Resultados das predições utilizando RMSE do cenário 2 para o conjunto de teste.

Vitimização por *bullying*

Variável	***Linear Regression***	***Decision Tree***	***Random Forest***	***XGBoost***
	RMSE	**RMSE**	**RMSE**	**RMSE**
Todas as variáveis	0.63	0.78	**0.59***	0.68
Caráter pessoal (cor/raça, gênero, auto-imagem e tipo de escola	**0.61***	0.80	0.63	0.75
Relação com amigos e professores	0.69	0.81	**0.60***	0.76
Sono (demora para dormir, quantidade de sonhos ruins e qualidade do sono	0.61	0.89	**0.60***	0.71

*Melhor resultado (menor RMSE) entre os modelos para um mesmo conjunto de variáveis.

Notamos que no cenário 2 as coisas não mudam muito. Na Tabela 4, podemos observar que ao considerar todas as variáveis e os conjuntos de variáveis de relação e sono, o modelo que também apresentou o menor erro (RMSE) para o teste foi o *Random Forest*, com um RMSE de 0.59, 0.60 e 0.60, respectivamente. Foi mantido os mesmos parâmetros considerados do cenário 1. Dentre os três melhores cenários, o conjunto com todas as variáveis foi o destaque deste cenário, com RMSE de 0.59.

No geral, nenhum RMSE é extremamente baixo (o que seria o ideal), mas é possível perceber que o modelo *Random Forest* foi o que mais se destacou, dentre os demais, com o menor RMSE deste cenário. O modelo *Linear Regression* também merece menção honrosa, detentor, nos dois cenários, de uma boa classificação. Porém, regressores baseados em árvores se mostram relativamente satisfatórios.

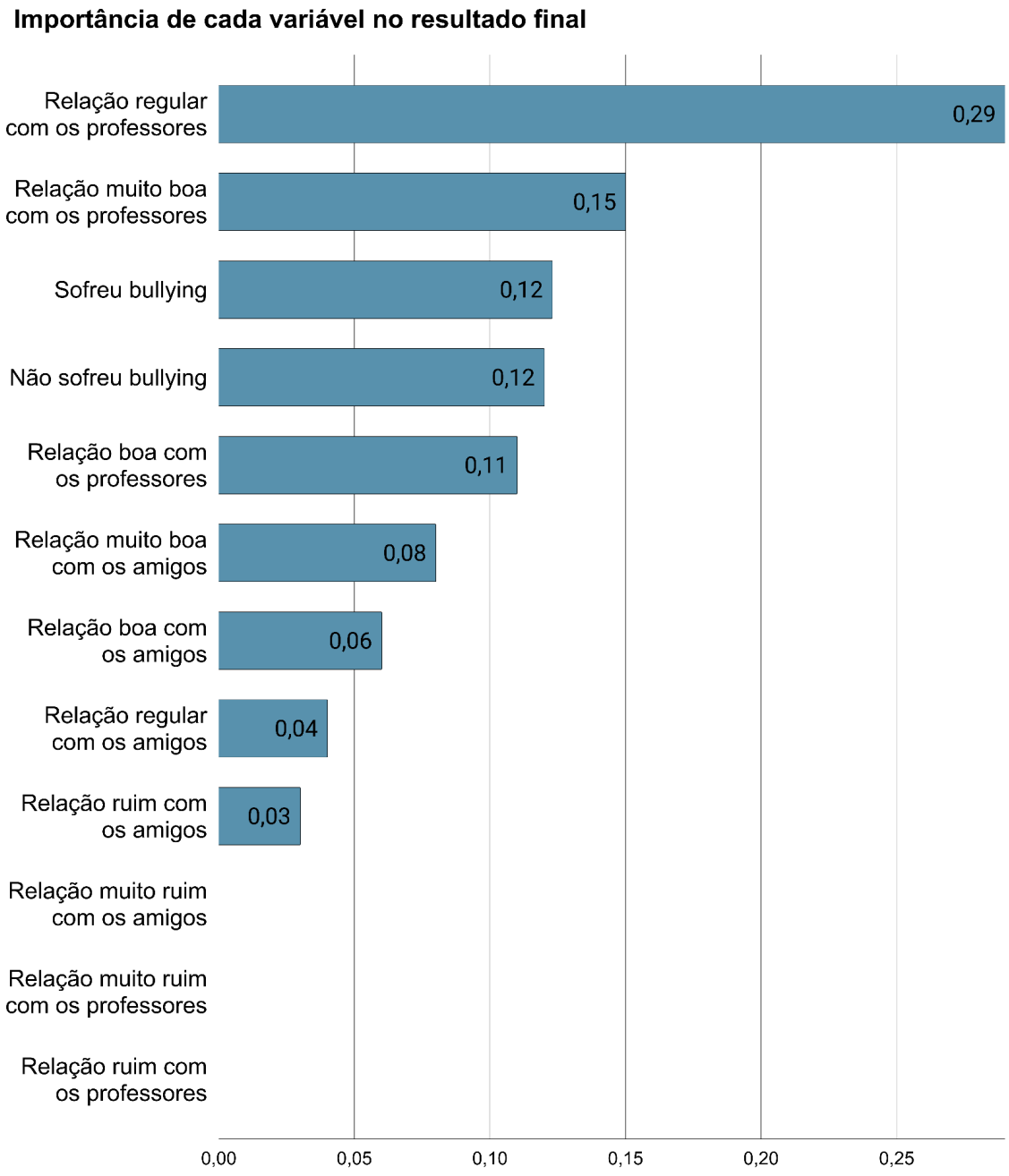
7.3 Interpretações dos modelos preditivos

A partir dos modelos que mais se destacaram e do pacote do *Python*, ELI5, que auxilia na visualização, apresentaremos as variáveis que mais influenciaram os resultados da predição, ou seja, quais foram as variáveis mais importantes/relevantes. Através da função *show_weights*, temos o peso associado a cada característica. O valor diz quanto impacto um recurso tem nas predições em média. A fim de facilitar a visualização, essas variáveis serão apresentadas em um gráfico de barras, onde o eixo vertical é representado pelas variáveis, e no eixo horizontal, o peso da importância da variável. Quanto maior o peso atribuído, mais importante é para o modelo.

7.3.1 Melhor cenário

Dentre os dois cenários apresentados, concluímos que o cenário 1, que considera a variável binária se sofreu ou não *bullying*, é o que detém o melhor modelo com o menor RMSE, sendo ele o regressor *Random Forest* com o conjunto de variáveis de relação com amigos e professores. Seu RMSE foi 0.57, o melhor apresentado dentre os cenários abordados. Baseado neste modelo, iremos avaliar quais foram as características mais importantes para a predição do modelo. As variáveis presentes estão descritas na seção 6.2.

Figura 7 - Importância das relações com amigos e professores na predição do cenário 1



Fonte: Elaborado pelo autor

Conforme a Figura 7, com o conjunto específico de variáveis de relações (amigos e professores), é possível observar que ter uma relação regular com os professores foi a característica mais importante para o modelo usando o algoritmo *Random Forest*. O peso é relativamente alto, onde esse foi o atributo com maior efeito no modelo para prever as notas.

Para os autores Kibriya, Xu e Zhang (2015 apud OLIVEIRA, et al., 2016, p. 5), o efeito do *bullying* diminui para os alunos que possuem a presença de uma professora em sala de aula, visto que são os profissionais com um contato mais próximo aos alunos dentro das escolas. Ainda segundo Wang, Brinkworth & Eccles (2013), a relação professor-aluno é crítica para o desenvolvimento saudável e para o processo acadêmico dos alunos, uma vez que o professor é a principal fonte de ligação entre alunos e educação. Assim, tanto para este trabalho como para a literatura especializada, é importante que o professor tenha um bom relacionamento com os alunos.

8 CONCLUSÃO E TRABALHOS FUTUROS

O objetivo deste trabalho foi o estudo e a criação de modelos regressores de aprendizado de máquina para a predição de notas escolares de alunos do ensino médio e técnico integrado a partir de indicadores escolares, incluindo indicadores de *bullying*. A partir dos resultados dos modelos preditivos, o melhor algoritmo classificado foi o *Random Forest*, com a menor raiz quadrática média de erros (0.57) em ambos os cenários avaliados no conjunto de variáveis de relação com amigos e professores. De acordo com esse algoritmo, foi possível ter noção de quais foram as características mais importantes para predição das notas dos estudantes, tornando capaz a melhoria e a interpretabilidade do modelo, podendo reduzir sua dimensionalidade (por manter aquelas que tiveram maior peso e excluir as com pesos baixos) e determinar, assim, quais características atribuem mais ao poder preditivo ao modelo. Dessa forma, concluímos que ter uma relação regular com os professores foi a variável que mais influenciou nas notas dos alunos.

Para trabalhos futuros, é desejável aprofundar mais nesses indicadores através de técnicas de estatísticas matemáticas, como, por exemplo, análise multivariada e/ou testes de hipóteses, que serão capazes de validar estatisticamente se os dados amostrais trazem evidências que apoiem ou não uma hipótese formulada, tornando ainda mais possível uma comparação direta com a literatura especializada.

REFERÊNCIAS

ALZUBI, J; NAYYAR, A; KUMAR, A. **Machine Learning from Theory to Algorithms: An Overview**. Disponível em: <https://www.researchgate.net/publication/329329261_Machine_Learning_from_Theory_to_Algorithms_An_Overview>. Acesso em: 23 abr. 2023.

CÉSAR, P. **Pré-Processamento de Dados | Conheça as Técnicas e as Etapas.** 2019. Disponível em: <https://www.datageeks.com.br/pre-processamento-de-dados/#:~:text=Existem%203%20principais%20passos%20envolvidos,dados%20e%20redu%C3%A7%C3%A3o%20de%20dados>. Acesso em: 22 abr. 2023.

CÉSAR, P. **XGBoost | Algoritmo de Machine Learning que está dominando!** 2019. Disponível em: <https://www.datageeks.com.br/xgboost/>. Acesso em: 23 abr. 2023.

COSTA, P.; PEREIRA, B. **O bullying na escola: a prevalência e o sucesso escolar.** 2010. I Seminário internacional "Contributos da Psicologia em Contextos Educativos", Braga: Universidade do Minho. Disponível em: <https://repositorium.sdum.uminho.pt/bitstream/1822/13613/1/Bullying%20na%20escola%20A%20prevalencia%20e%20o%20sucesso%20escolar.pdf>. Acesso em: 01 out 2022.

COSTA, R. **Entenda o que é a Regressão Linear**. 2019. Disponível em <https://www.linkedin.com/pulse/entendendo-teoria-da-regressão-linear-ruan-costa/?originalSubdomain=pt>. Acesso em: 23 abri. 2023.

DADOS, Ciência e. **Por que e como data science é mais do que apenas machine learning?** 2021. Disponível em: <https://www.cienciaedados.com/por-que-e-como-data-science-e-mais-do-que-apenas-machine-learning/>. Acesso em: 09 abril 2022.

DAMACENO, L. **Regressão Linear?** 2020. Disponível em: <https://medium.com/@lauradamaceno/regress%C3%A3o-linear-6a7f247c3e29>. Acesso em: 23 abr. 2023.

DIDÁTICA, T. **Como funciona o algoritmo de Árvore de Decisão**. 2020. Didática Tech. Disponível em: <https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/>. Acesso em: 23 abr. 2023.

DIDÁTICA, T. **Entenda o que são Dados de Treino e Teste (Machine Learning)**. 2019. Disponível em: <https://didatica.tech/dados-de-treino-e-teste/#:~:text=Uma%20das%20bases%20do%20machine,usados%2C%20melhor%20ficar%C3%A1%20o%20modelo.>. Acesso em: 31 maio. 2023.

ESCOVEDO, T. **Machine Learning: Conceitos e Modelos - Parte I: Aprendizado Supervisionado**. 2020. Disponível em: <https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445>. Acesso em: 23 abr. 2023.

ESCOVEDO, T. **Implementando um Modelo de Classificação no Scikit-Learn***. 2021. Disponível em: <https://tatianaesc.medium.com/implementando-um-modelo-de-classifica%C3%A7%C3%A3o-no-scikit-learn-6206d684b377>. Acesso em: 23 abr. 2023.

GÉRON, A. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**. Amazon.com.br, p. 9. 2019. Disponível em: <https://www.amazon.com.br/M%C3%A3os-Obra-Aprendizado-Scikit-Learn-TensorFlow/dp/8550803812>. Acesso em: 31 maio. 2023.

HELENA, M. **Aprenda o que são métodos de estimação**. Data Hackers. 2019. Disponível em: <https://medium.com/data-hackers/aprenda-o-que-são-métodos-de-estimação-ea48b189c039>. Acesso em: 08 maio 2022.

IBGE. **Pesquisa nacional de saúde escolar**. IBGE, p. 40. 2019. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101852>. Acesso em: 09 abril 2022.

NANSEL et al. **Bullying Behaviors Among US Youth: Prevalence and Association With Psychosocial Adjustment**. JAMA Psychiatry. 2001. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435211/>. Acesso em: 09 abril 2022.

NETO, A. A. L. **Bullying: comportamento agressivo entre estudantes**. Jornal de Pediatria. 2005. Disponível em: <https://doi.org/10.1590/S0021-75572005000700006>. Acesso em: 09 abril 2022.

MIRANDA, J. **get_dummies vs OneHotEncoder: qual método escolher?** 2022. Disponível em: <https://www.alura.com.br/artigos/get-dummies-vs-onehotencoder-qual-metodo-escolher#:~:text=Ao%20aplicar%20em%20novos%20dados,em%20modelos%20de%20machine%20learning.>. Acesso em: 22 abr. 2023.

OLIVEIRA, B. **Qual a relação entre machine learning e a estatística?** OpenData. 2020. Disponível em: <https://operdata.com.br/blog/a-relacao-entre-machine-learning-e-a-estatistica/>. Acesso em: 09 abril 2022.

OLIVEIRA, et al. **Bullying e violência no ambiente escolar: uma revisão de literatura no período de 2015-2019**. Revista Eletrônica Acervo Saúde, 11(13), e860. 2019. Disponível em: <https://doi.org/10.25248/reas.e860.2019>. Acesso em: 09 abril 2022.

OLIVEIRA, et al. **O efeito do bullying no desempenho dos alunos: Uma análise para as escolas públicas do Recife**. 2016. Disponível em: <*https://docplayer.com.br/56355967-Efeito-do-bullying-no-desempenho-dos-alunos-uma-analise-para-as-escolas-publicas-do-recife.html*>. Acesso em: 09 abril 2022.

PIGOZI, P. L; MACHADO, A. L. **Bullying na adolescência: visão panorâmica no Brasil.** Ciência & Saúde Coletiva. 2015. Disponível em: <https://doi.org/10.1590/1413-812320152011.05292014>. Acesso em: 10 abril 2022.

PINHEIRO, N. **Pré-processamento de dados com Python**. 2021. Data Hackers. Disponível em: <https://medium.com/data-hackers/pré-processamento-de-dados-com-python-53b95bcf5ff4>. Acesso em: 22 abr. 2023.

PONZO, M. **Does bullying reduce educational achievement? An evaluation using matching estimators**. 2012. Munich Personal RePEc Archive. Disponível em: <https://mpra.ub.uni-muenchen.de/36064/>. Acesso em: 01 out 2022.

QUINTANILHA, C. **Um olhar exploratório sobre a percepção do professor em relação ao fenômeno bullying**. 2011. UERJ.

RAFAEL. **Guia Básico de Pré-Processamento de Dados**. 2020. Disponível em: <https://sigmoidal.ai/guia-basico-de-pre-processamento-de-dados/#:~:text=Standard%20Scaler&text=Padroniza%20as%20features%20removendo%20a,mais%20manej%C3%A1veis%20para%20nossos%20modelos.>. Acesso em: 23 abr. 2023.

RIZZOTTO, J.S; FRANÇA, M. T. A. **O bullying afeta o desempenho escolar dos alunos brasileiros? Uma análise por meio do PISA 2015**. ABER. 2021. Disponível em: <https://brsa.org.br/wp-content/uploads/wpcf7-submissions/1066/Identificado-O-bullying-afeta-o-desempenho-escolar-dos-alunos-Uma-análise-através-do-PISA-2015.pdf>. Acesso em: 09 abril 2022.

SILVIA. **Tipos de variáveis**. 2012. Disponível em: <http://leg.ufpr.br/~silvia/CE055/node8.html>. Acesso em: 22 abril 2023.

SILVA, M. **Bullying: sua origem e evolução**. Aprendizagem. 2015. Disponível em: <https://www.mouracoaching.com/origem-e-evolucao-do-bullying/>. Acesso em: 09 abril 2022.

SILVA et al. **O olhar de professores sobre o bullying e implicações para a atuação da enfermagem**. Faculdade de Enfermagem. 2014. Disponível em: <http://repositorio.bc.ufg.br/handle/ri/17440>. Acesso em: 09 abril 2022.

SILVEIRA, G.; BULLOCK, B. **Machine Learning - Introdução à Classificação**. 2017. Disponível em: <https://www.casadocodigo.com.br/products/livro-machine-learning>. Acesso em: 09 abril 2022.

SMITH, P; MADSEN, K; MOODY, J. **What causes the age decline in reports of being bullied at school? Towards a developmental analysis of risks of being bullied**. 1999.

Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/0013188990410303>. Acesso em: 7 maio. 2023.

SOUSA, G. **Entendendo os Modelos de Regressão**. 2022. Disponível em: <https://www.learningdata.dev/post/entendendo-os-modelos-de-regressao#viewer-fj6hv>. Acesso em: 23 abr. 2023.

SOUZA, G.; ORELLANA, V.; LEIVAS, P. **Impacto do Bullying no Atraso Escolar**. 2019. ResearchGate. Disponível em <https://www.researchgate.net/publication/339697010_IMPACTO_DO_BULLYING_NA_PERFOMANCE_ESCOLAR>. Acesso em: 01 out 2022.

SCUDILIO, J. **Parte III: Como utilizar modelos de Machine Learning para reduzir o Churn**. 2020. Disponível em: <https://www.flai.com.br/juscudilio/parte-iii-como-utilizar-modelos-de-machine-learning-para-reduzir-o-churn/>. Acesso em: 23 abr. 2023.

VASCONCELLOS, P. **Como saber se seu modelo de Machine Learning está funcionando mesmo**. Disponível em: <https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-est%C3%A1-funcionando-mesmo-a5892f6468b>. Acesso em: 23 abr. 2023.

VIEIRA-JUNIOR, F. U; VIEIRA, K. M. R; MORETTI, A. C. **Bullying com adolescentes escolares em diferentes contextos educacionais.** Revista de Enfermagem. 2020. Disponível em: <https://periodicos.ufpe.br/revistas/revistaenfermagem/article/view/243622> Acesso em: 10 abril 2022.

VIGNOLES, A.; MESCHI, E. **The Determinants of Non-Cognitive and Cognitive Schooling Outcomes. Report to the Department of Children, Schools and Families**. 2010. Centre for the Economics of Education. Disponível em: <https://eric.ed.gov/?id=ED529936>. Acesso em: 01 out 2022.

WANG, M.-T., BRINKWORTH, M., & ECCLES, J. **Moderating effects of teacher–student relationship in adolescent trajectories of emotional and behavioral adjustment**. 2013. Developmental Psychology. Disponível em: <https://doi.org/10.1037/a0027916>. Acesso em: 08 mai 2023.

ZEFERINO, D. **Dados, informação e conhecimento: qual a diferença dos conceitos?** Certifiquei. 2020. Disponível em: <https://www.certifiquei.com.br/dados-informacao-conhecimento/>. Acesso em: 14 maio 2022.