Adriano Malerba

Previsão de Evasão Universitária através de Aprendizado de Máquina

Brasil

Julho de 2023

Dedico este trabalho aos meus pais, Wilson Malerba e Maria Auxiliadora Malerba e a minha linda esposa Ericka Priscilla Marques Malerba e aos meus filhos Arthur, Alanna e Álefe.

Agradecimentos

Agradeço, primeiramente, a Deus pelo dom da vida e, principalmente, por se fazer presente em todos os momentos de minha vida, além de me dar sabedoria e forças para vencer mais uma etapa. “Deus não escolhe os capacitados. Ele capacita os escolhidos.”

A minha esposa Ericka Priscilla Marques Malerba pela demonstração de amor se tornando peça fundamental para o desenvolvimento da pesquisa e, consequentemente, co-autor deste trabalho.

Aos meus familiares, pelo apoio e compreensão.

Agradeço imensamente ao Professor Carlos Henrique Valério de Moraes por ter depositado sua confiança em mim; pela disposição e paciência em orientar o desenvolvimento do trabalho.

Agradeço também a todas as pessoas que, de forma direta ou indireta, colaboraram para a elaboração desta dissertação, sinceramente, meus agradecimentos.

Resumo

A evasão de alunos é um problema que afeta as instituições de ensino superior no mundo todo, tendo impactos negativos tanto para os alunos quanto para as instituições, sejam elas públicas ou privadas. É essencial, que as instituições tenham ferramentas que os auxiliem no controle da evasão, proporcionando aos gestores a compreensão das expectativas educacionais dos alunos que ingressam no ensino superior, a fim de aprimorar a compreensão desse fenômeno. Os estudos recentes sobre previsão de evasão universitária utilizando aprendizado de máquina representa um avanço significativo na área da educação. Ao empregar a Técnica de Validação Cruzada (K-Fold) juntamente com uma variedade de algoritmos de classificação, como árvores de decisão, regressão logística, floresta aletória e máquinas de vetores de suporte, entre outros. Este trabalho busca entender e antecipar os padrões de evasão entre os alunos. Essa abordagem inovadora não apenas identifica fatores de risco para a evasão, mas também fornece informações valiosas para instituições educacionais no desenvolvimento de estratégias proativas de retenção de alunos. Ao prever com precisão a probabilidade de um estudante abandonar seus estudos, as universidades podem intervir precocemente, oferecendo suporte personalizado e recursos adicionais para ajudar os alunos a superar desafios acadêmicos e pessoais.Para isso, em relação ao modelo de recuperação dos alunos as técnicas LogisticRegression, GradientBoosting e XG Boost obtiveram resultados semelhantes e promissores, acima de 90% para F1-Score de formando e F1-score de evasão próximo a 89%. Já para os casos de algoritmos interpretáveis, modelo para desligamento de Alunos, os melhores resultados foram para os modelos Random Forest e Decision Tree com valores de 91% para F1-Score de Formando, 84% para F1-score de evasão. Este trabalho representa uma contribuição significativa para a melhoria da qualidade e da eficácia dos programas educacionais, promovendo a retenção e o sucesso dos alunos universitários.

Palavras-chave: Evasão de alunos, aprendizado de máquina, previsão, seleção de modelos.

Abstract

Student dropout is a problem that affects higher education institutions around the world as a whole, having negative impacts on both students and institutions, whether whether public or private. It is essential that institutions have tools that help control evasion, providing managers with an understanding of expectations educational outcomes for students entering higher education, in order to improve their understanding of this phenomenon. Recent studies on university dropout prediction using Machine learning represents a significant advance in the field of education. To the employ the Cross Validation Technique (K-Fold) along with a variety of classification algorithms such as decision trees, logistic regression, random forest and support vector machines, among others. This work seeks to understand and anticipate dropout patterns among students. This innovative approach not only identifies risk factors for evasion, but also provides valuable information for institutions educational institutions in developing proactive student retention strategies. When predicting accurately determine the probability of a student abandoning their studies, universities can intervene early, offering personalized support and additional resources to help students overcome academic and personal challenges. To this end, in relation to the model of student recovery using the LogisticRegression, GradientBoosting and XG Boost techniques obtained similar and promising results, above 90% for F1-Score of graduates and F1-evasion score close to 89%. For cases of interpretable algorithms, model for student dismissal, the best results were for the Random models Forest and Decision Tree with values of 91% for Graduate F1-Score, 84% for F1-score of evasion. This work represents a significant contribution to improving the quality and effectiveness of educational programs, promoting retention and success of university students.

Keywords: Student dropout, machine learning, prediction, model selection.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	ABsurdas Normas para TeX
AD	Árvore de Decisão
DT	Decision Tree
ENEM	Exame Nacional do Ensino Médio
FIES	Fundo de Financiamento Estudantil
IA	Inteligência Artificial
IES	Instituição de Ensino Superior
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IBGE	Instituto Brasileiro de Geografia e Estatística
KNN	K-Nearest Neighbors
ML	Machine Learning
ProUni	Programa Universidade para Todos
PNAD	Pesquisa Nacional por Amostra de Domicílios
RNA	Redes Neurais Artificiais
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SVM	Support Vector Machine

Sumário

1 Introdução

A evasão universitária é um problema que afeta diversas instituições de ensino superior no mundo todo. Este fenômeno pode ser definido como a interrupção do curso por parte do aluno antes da conclusão do mesmo, podendo ocorrer por diversos motivos, como problemas financeiros, desmotivação ou falta de preparo acadêmico. A evasão escolar tem impactos negativos tanto para os alunos quanto para as instituições de ensino, incluindo perda de tempo, recursos e investimentos (FILHO et al., 2007).

Essas discussões são relevantes no contexto atual, considerando a expansão das políticas de acesso e permanência na educação superior. É essencial acompanhar de perto os estudantes que entram no sistema para garantir a eficácia dessas políticas. A atuação das instituições desempenha um papel estratégico no controle da evasão, exigindo que os gestores compreendam as expectativas educacionais dos alunos que ingressam no ensino superior, a fim de aprimorar a compreensão desse fenômeno (TINTO, 1975).

Nesse contexto, o uso de técnicas de Aprendizado de Máquina, em inglês *machine learning*, tem se mostrado uma ferramenta eficaz para prever a evasão escolar nos cursos de graduação. Afinal, o aprendizado de máquina é uma área da inteligência artificial que utiliza algoritmos e modelos estatísticos para analisar dados e fazer previsões. Fazendo-se uso desta ferramenta é possível identificar os alunos em risco de evasão e tomar medidas preventivas para ajudá-los a permanecer no curso (FRANK; HALL; WITTEN, 2016).

Este trabalho tem como objetivo aplicar a técnica de Aprendizado de Máquina como ferramenta para prever a evasão escolar nos cursos de graduação. Para isso, será utilizada a técnica de validação cruzada (k-fold) para avaliar a capacidade de generalização do modelo em uma base de dados contendo informações sobre alunos que abandonaram o curso e alunos que permaneceram no curso. Com base nesses dados, serão aplicados algoritmos para prever a evasão escolar. Este trabalho visa contribuir para a área de educação ao fornecer um modelo eficaz para prever a evasão escolar na graduação. O que pode ajudar as instituições de ensino a identificar os alunos em risco e tomar medidas preventivas para evitá-la.

1.1 Justificativa

Nos anos 90, o Brasil enfrentou um aumento significativo na taxa de evasão escolar na graduação, que foi atribuído a diversos fatores, como a falta de preparo dos estudantes, a falta de apoio financeiro e as condições precárias de ensino. Na época, o governo federal adotou algumas medidas para tentar reduzir a evasão, como a criação de programas

de bolsas de estudo e a ampliação do acesso à educação superior por meio do ProUni (Programa Universidade para Todos) e do FIES (Fundo de Financiamento Estudantil) (BRASIL, 2015).

A partir dos anos 2000, o país registrou uma queda na taxa de evasão, que foi atribuída a diversos fatores, como a melhoria da qualidade do ensino, o aumento do número de vagas nas universidades e a adoção de políticas públicas para apoiar os estudantes. No entanto, mesmo com a redução, a taxa de evasão no Brasil ainda é considerada alta, especialmente em cursos noturnos e em instituições de ensino superior privadas (Brasil, 2015). Embora as taxas de evasão variem entre as instituições e regiões, elas podem ter um impacto significativo na qualidade da educação, bem como na sustentabilidade financeira das instituições de ensino. Além disso, a evasão de alunos também pode afetar negativamente o desenvolvimento pessoal e profissional dos estudantes, bem como sua capacidade de contribuir positivamente para a sociedade (TINTO, 1975; TINTO, 1993; PASCARELLA; TERENZINI, 1991).

Raramente uma instituição possui uma plataforma capaz de converter grandes volumes de dados de diversas áreas em conhecimento específico para os gestores escolares. A coleta de informações ainda é lenta, parcial e suscetível a erros, o que sobrecarrega o sistema, aumenta os chamados ao setor de informática e não oferece suporte rápido, adequado e eficaz à tomada de decisão (Souza, 2020). A aprendizagem de máquina (ML) é um campo da ciência da computação que se originou do estudo do reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial (IA). Ele se concentra em capacitar as máquinas a realizar tarefas que normalmente requerem intervenção humana. A ML envolve a programação de computadores com regras predefinidas que permitem que tomem decisões com base nos dados disponíveis (BIAMONTE et al., 2017).

Nos últimos anos, as instituições de ensino superior têm investido em estratégias para combater a evasão, como a oferta de tutorias, a criação de programas de apoio psicológico e financeiro aos estudantes, e o uso de tecnologias educacionais para melhorar a aprendizagem. Além disso, tem havido uma maior atenção para a implementação de modelos de predição de evasão, como o uso de técnicas de ML, com o objetivo de identificar precocemente os alunos em risco de evasão e adotar medidas preventivas (TEIXEIRA; MENTGES; KAMPFF, 2019).

Através desta pesquisa será possível analisar por meio de aprendizado de máquina as principais variáveis de fatores que causam a evasão de alunos de graduação e identificar estratégias eficazes para prevenir e combater esse problema. Espera-se que os resultados desta pesquisa contribuam para o desenvolvimento de políticas e práticas educacionais mais eficazes, bem como para a compreensão mais aprofundada dos desafios enfrentados pelos estudantes de graduação e pelas instituições de ensino superior;

1.2 Objetivos

Nesta seção, são apresentados os objetivos, geral e específicos, que orientaram a construção dessa pesquisa.

1.2.1 Objetivo Geral

Desenvolver e avaliar um modelo de aprendizado de máquina para predição de evasão de alunos de graduação, com o intuito de apoiar a tomada de decisão institucional para prevenir a evasão e promover a retenção dos alunos.

1.2.2 Objetivos específicos

A fim de alcançar o objetivo geral deste trabalho, o mesmo foi desdobrado em alguns objetivos específicos:

1. Buscar na bibliografia existente as possíveis causas da evasão escolar e os modelos e teorias relacionados a essa questão;

2. Coletar e processar dados dos alunos e do ambiente acadêmico;

3. Treinar diferentes modelos de aprendizado de máquina que possam identificar os alunos em risco de evasão;

4. Testar e validar os modelos;

5. Analisar os resultados obtidos.

1.2.3 Estrutura do Trabalho

O presente trabalho encontra-se estruturado em 5 (cinco) capítulos.

O primeiro capítulo traz a contextualização e justificativa do tema, bem como os objetivos gerais e específicos do trabalho e a sua estrutura.

No segundo capítulo é apresentado o fundamentação teórica que norteou o estudo.

O terceiro capítulo, é apresentado e discutido outros trabalhos acadêmicos ou pesquisas relevantes que abordam temas similares ou relacionados a este estudo.

O quarto capítulo descreve a metodologia adotada, tratando da classificação da pesquisa.

No quinto capítulo são apresentados os resultados do trabalho, de forma clara e objetiva, utilizando tabelas, gráficos para facilitar a compreensão.

Por fim, no capítulo sexto são apresentadas as conclusões obtidas através da pesquisa, sugestões de possibilidades de trabalhos futuros, seguidas das referências biblio-

gráficas.

O trabalho é complementado com o link do código fonte usado na execução da metodologia do trabalho.

2 Fundamentação Teórica

Nesta seção, serão abordadas teorias, conceitos e estudos anteriores que sustentam a pesquisa. Ao analisar o conhecimento existente, busca-se contextualizar o problema de pesquisa e contribuir para o avanço do campo. A fundamentação teórica permite identificar lacunas no conhecimento e destacar a relevância e originalidade do estudo proposto.

2.1 Evasão Escolar

O conceito de evasão escolar tem sido amplamente estudado devido às diversas interpretações sobre o tema. Essa diversidade de conceitos dificulta a análise das causas e a proposição de soluções para um problema que persiste desde o início da educação formal até os dias atuais (SILVA, 2022).

Para (FIALHO et al., 2014), a evasão escolar pode ser conceituada como a interrupção do processo educacional pelo estudante, em qualquer modalidade de ensino, que o impeça de alcançar um nível de compreensão adequado para sua formação.

Segundo (FILHO et al., 2007), a evasão de alunos é um fenômeno complexo que envolve uma série de fatores, incluindo características individuais dos alunos, características institucionais e fatores externos, como a situação econômica do país.

De acordo com (RIFFEL; MALACARNE, 2010), a evasão é o ato de evadir-se, fugir, abandonar, sair, desistir, não permanecer em algum lugar. Quando se trata de evasão escolar, entende-se a fuga ou abandono da escola em função da realização de outra atividade.

Para (BISSOLI; RODRIGUES, 2017), refere-se à situação em que o aluno abandona a escola durante o ano letivo. As causas dessa decisão estão frequentemente relacionadas a problemas familiares, a necessidade de trabalhar para contribuir com o orçamento familiar, falta de interesse nos estudos, dificuldade em compreender o conteúdo ensinado pelos professores, entre outros fatores.

De acordo com (ASTIN, 1999), a evasão de alunos pode ser entendida como uma consequência da falta de envolvimento dos alunos com a instituição e com as atividades acadêmicas. Para o autor, o envolvimento dos alunos é fundamental para a retenção e o sucesso acadêmico.

(LOPES, 2023) complementa essa definição, afirmando que a evasão também pode ocorrer quando os estudantes "transferem-se para outras instituições de ensino ou mudam de curso dentro da mesma universidade". Nesse sentido, a evasão não está restrita apenas

ao abandono definitivo da instituição, mas também inclui a interrupção da trajetória acadêmica inicialmente escolhida.

(TINTO, 1993), define a evasão de alunos como "a saída prematura e voluntária dos estudantes de um curso ou programa educacional antes de sua conclusão". Essa definição enfatiza o caráter voluntário do abandono dos estudos por parte dos alunos. O autor apresentou um modelo revisado que inclui dimensões inter-relacionadas para avaliar o nível de integração dos estudantes na vida acadêmica. A dimensão Experiências Institucionais inclui o desempenho acadêmico, interações com professores e funcionários, atividades extracurriculares e interações com colegas como fatores determinantes para a integração acadêmica e social dos estudantes.

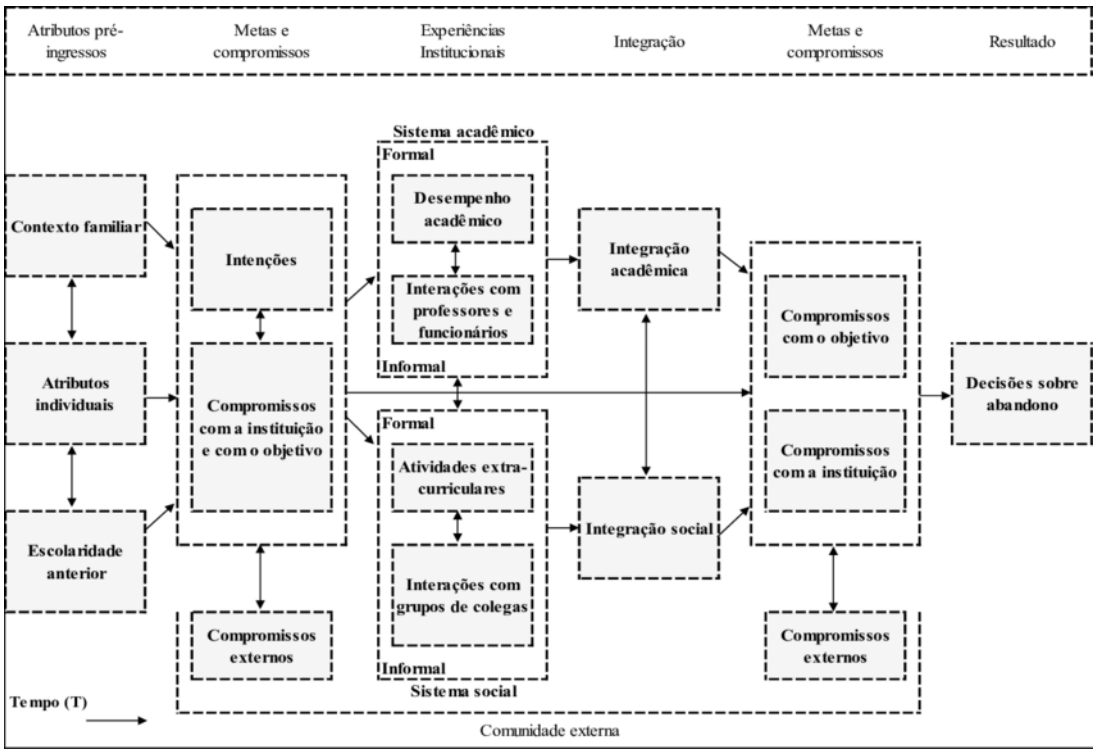

Figura 1 – Modelo de abandono institucional

Fonte: Tinto (1993)

De acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 1998), a evasão escolar é caracterizada pela saída definitiva do aluno do sistema educacional, ocorrendo antes da conclusão do ano, série ou ciclo, por desistência, independentemente do motivo. Isso representa um fracasso em relação ao objetivo de promover o aluno para um nível de ensino superior, impactando o desenvolvimento cognitivo, habilidades e competências esperadas para aquele nível de ensino.

Segundo o Índice de Desenvolvimento da Educação Básica - (IDEB, 2007), abandono escolar é o ato de afastar-se do sistema educacional e desistir das atividades escolares sem solicitar transferência.

Há falta de consenso na literatura internacional sobre a definição de evasão escolar, especialmente no nível universitário. (TINTO, 1975), um dos clássicos no assunto, define abandono escolar como a saída do aluno da Instituição de Ensino Superior (IES) sem receber o diploma. Essa definição, amplamente reconhecida internacionalmente, foi adotada no Brasil pela Comissão Especial de Evasão do Ensino Superior em 1996 e ainda é amplamente utilizada para calcular as taxas de evasão (PRESTES; FIALHO, 2018).

A partir de 2020, o Mapa do Ensino Superior no Brasil passou a ser produzido pela equipe do Instituto Semesp, um centro de inteligência analítica criado pelo Semesp com o objetivo de compartilhar para pesquisadores, educadores, gestores privados e públicos, jornalistas e para a sociedade em geral informações relevantes e confiáveis que lhes permitam tomar decisões, estabelecer estratégias ou formular políticas públicas, visando o desenvolvimento da educação superior. A equipe do Instituto Semesp usou como guia para a elaboração do Mapa do Ensino Superior no Brasil 2023 os dados do Censo da Educação, referentes a 2021, divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) em novembro de 2022, e outras fontes como IBGE, microdados do ENEM e do PROUNI, CAGED, RAIS, Big Data Analytics, entre outros (SEMESP, 2023).

Na figura 2 são apresentados os percentuais de evasão no Ensino Superior no Brasil das Redes Públicas e Privadas, no cursos presenciais e a distância, de 2014 a 2019.

Na figura 3 são apresentadas as taxas de evasão na Rede Privada, calculadas para os 20 cursos com maior número de alunos em 2019 nas instituições privadas de ensino superior no Brasil.

Na figura 4 são apresentadas as taxas de evasão no curso no 1º. Ano, nos cursos de graduação presenciais.

Na figura 5 são apresentados os percentuais de evasão no Ensino Superior no Brasil das Redes Públicas e Privadas, até o ano de 2021.

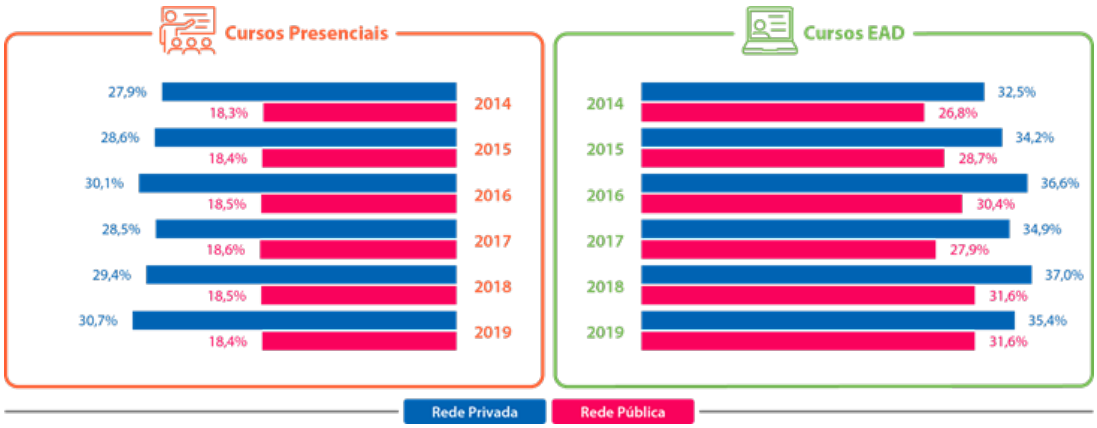

Figura 2 – Evasão no Ensino superior no Brasil até 2019

Fonte: (SEMESP, 2021)

Cursos Presenciais

Curso	Taxa de Evasão
Sistemas de Informação	37,6%
Administração	35,9%
Educação Física	34,3%
Engenharia Mecânica	34,2%
Engenharia de Produção	33,5%
Publicidade e Propaganda	33,0%
Contabilidade	32,9%
Engenharia Civil	31,5%
Nutrição	31,4%
Biomedicina	30,6%
Enfermagem	29,9%
Fisioterapia	29,1%
Arquitetura e Urbanismo	28,4%
Pedagogia	27,9%
Direito	27,6%
Psicologia	27,1%
Farmácia	24,1%
Medicina Veterinária	23,4%
Odontologia	19,0%
Medicina	6,8%

Cursos EAD

Curso	Taxa de Evasão
Marketing	44,7%
Matemática Formação de Professor	44,3%
Letras Português Formação de Professor	44,1%
Gestão Comercial	42,5%
História Formação de Professor	42,0%
Gestão Financeira	41,7%
Sistemas de Informação	41,3%
Logística	40,9%
Gestão Ambiental	40,5%
Gestão de Pessoas	38,6%
Gestão de Negócios	37,5%
Engenharia de Produção	37,2%
Gestão Pública	36,8%
Administração	36,5%
Contabilidade	35,3%
Serviço Social	34,7%
Educação Física Formação de Professor	31,6%
Enfermagem	30,6%
Educação Física	29,5%
Pedagogia	28,0%

Figura 3 – Taxa de Evasão na Rede Privada

Fonte: (SEMESP, 2021)

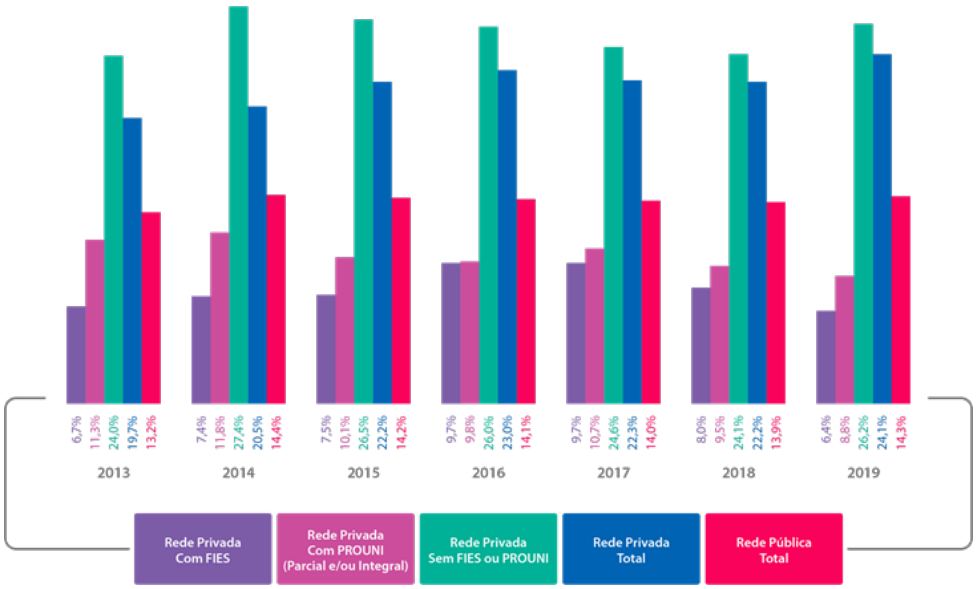

Figura 4 – Taxa de Evasão no Curso no 1º. Ano

Fonte: (SEMESP, 2021)

A tabela demonstra a importância que o financiamento tem para a escolha do curso. O aluno que ingressa com FIES entra mais vocacionado, escolhendo o curso e a IES que quer cursar, daí a menor evasão. Sem o FIES, a evasão é maior porque o estudante escolhe pela facilidade de ingresso e pelo preço do curso, sem levar em consideração a vocação (SEMESP, 2021).

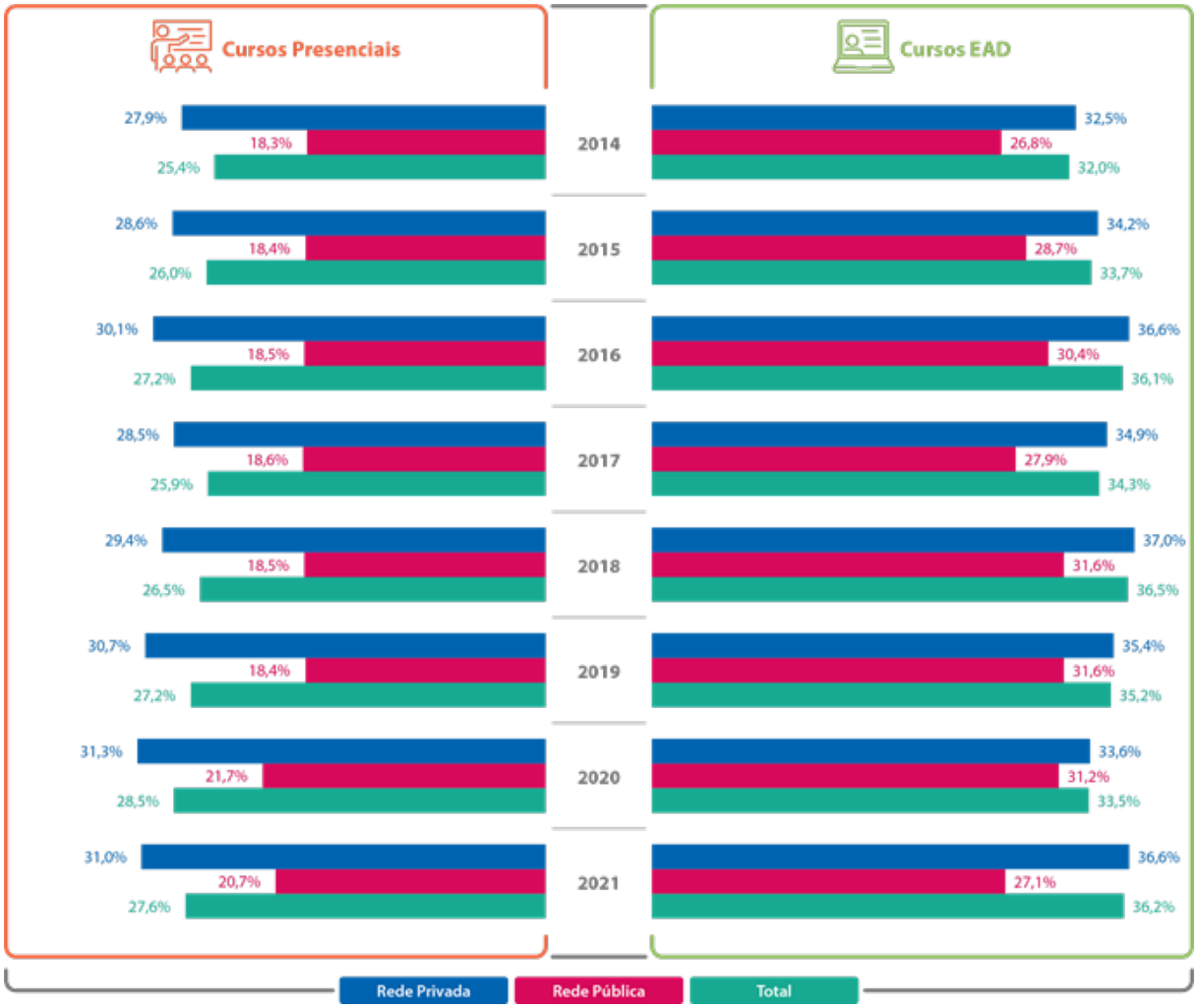

Figura 5 – Evasão no Ensino superior no Brasil

Fonte: (SEMESP, 2023)

2.1.1 Causas da Evasão de Alunos

As principais causas da evasão escolar no Brasil têm sido objeto de estudo e pesquisa por muitos anos, e há uma ampla literatura sobre o assunto.

(TINTO, 1975) destaca que nem toda evasão é igual, inicialmente distinguindo entre abandono involuntário e voluntário. O involuntário geralmente ocorre devido ao fracasso escolar. O abandono voluntário ocorre devido à falta de adaptação entre o estudante, o ambiente acadêmico e o sistema social da faculdade ou universidade.

O autor afirma ainda que as características individuais dos estudantes afetam a forma como eles se integram à instituição e influenciam sua probabilidade de permanecer ou desistir. Destacando:

a) O background familiar: o status socioeconômico, escolaridade, procedência, valores, expectativas e características de apoio;

b) Atributos individuais gerais: gênero, idade, etnia, habilidades, competências, capacidade de interação, características de personalidade;

c) A escolaridade anterior, incluindo desempenho acadêmico, talentos e experiências acadêmicas e sociais, afeta a percepção do indivíduo sobre sua própria competência e expectativa para o futuro, o que impacta seu compromisso em se graduar;

d) Os atributos motivacionais e expectativas em relação à formação acadêmica e carreira profissional são influenciados pela interação entre indivíduo, família e experiências educacionais prévias. Esses atributos são importantes preditores de evasão e consistem no compromisso com o objetivo de graduar-se e compromisso institucional, refletindo a avaliação do estudante sobre o ambiente universitário e consigo mesmo.

(SOUZA, 2020) em sua pesquisa sobre as possíveis causas de evasão apresentou 12 causas, apresentas na Tabela 1 a seguir:

Causas	Autores
Desmotivação	Coelho (2001), Frankola (2001), Neves (2006), Ramminger (2006)
Falta de companheiros presenciais	Frankola (2001), Neves (2006), Longo(2009)
Falta de tempo	Abraed (2008), Almeida (2008), Comarella (2009), Censo (2014), Neves (2006), Pacheco (2007), Ramminger (2006)
Falta de disciplina	Coelho (2001)
Problemas familiares	Almeida (2008), Ramminger (2006)
Questão financeira	Abraed (2008), Ramminger (2006)
Acham que curso seria mais fácil	Abraed (2008), Almeida (2008), Ramminger (2006)
Dificuldade de acesso ao computador e internet	Almeida (2008), Pacheco (2007)
Falta de preparo do professor	Sihler e Ferreira (2011)
Alta rotatividade de tutores	Almeida (2008)
Falta de feedback do tutor	Almeida (2008)
Falta de adaptação à EAD	Abraed (2008), Censo (2014), Longo (2009), Prensky (2001)

Tabela 1 – Causas de Evasão segundo alguns autores

Fonte: (SOUZA, 2020)

Segundo (TINTO, 1975), um dos principais teóricos sobre a evasão universitária, a falta de engajamento acadêmico é um dos principais fatores que levam os estudantes a abandonarem seus estudos. Esse engajamento pode ser influenciado por diversos fatores, como a qualidade do ensino, o relacionamento com os professores e colegas, e a satisfação com o curso.

De acordo com (BARBOSA, 2021), foram identificadas 14 causas do abandono escolar no Brasil, conforme mostrado na Tabela 2:

Item	Causas do Abandono	Descrição
1	ACESSO LIMITADO	O acesso à educação é um desafio em todo o Brasil, especialmente em áreas rurais e periferias urbanas, devido à escassez de escolas e transporte público. Para resolver isso, as políticas públicas devem otimizar a oferta de vagas, expandir a infraestrutura escolar, promover a educação à distância e melhorar o transporte escolar.
2	NECESSIDADE ESPECIAL	Cerca de 5% dos jovens abandonam a escola devido a limitações físicas. É fundamental que a escola se adapte às necessidades especiais, como previsto na legislação. Além disso, 37% dos jovens com necessidades especiais beneficiados pelo BPC não frequentam a escola, exigindo capacitação de professores e recursos multifuncionais.
3	GRAVIDEZ E MATERNIDADE	A maternidade precoce pode causar constrangimentos sociais e limitar o tempo disponível para os estudos. Uma pesquisa de 2016 realizada pelo MEC, OEI e Flacso revelou que 18% das meninas que pararam de estudar tiveram a gravidez como principal motivo.
4	ATIVIDADES ILEGAIS	O uso de drogas e atividades ilegais concorre com a frequência escolar. Ações preventivas e educativas, juntamente com medidas de repressão e controle do uso e comércio de drogas nas proximidades das escolas, são opções viáveis para resolver esse problema.
5	MERCADO DE TRABALHO	Um dos principais motivos para o afastamento dos jovens das atividades escolares é o envolvimento precoce e intenso com o mundo do trabalho. Jovens de 17 anos ou mais são os mais afetados por esse motivo.
6	POBREZA	Às vezes, os jovens não têm condições básicas de alimentação, vestuário ou higiene para frequentar a escola com dignidade, ou faltam estrutura em casa para realizar os deveres de casa. Existem programas que buscam atender essas necessidades e condicionam apoio financeiro à frequência escolar.
7	VIOLÊNCIA	As violências física e psicológica podem ocorrer em casa, na escola ou nas ruas, afetando negativamente o aprendizado dos jovens e desviando sua atenção dos estudos.
8	DÉFICIT DE APRENDIZAGEM	Reprovações repetidas são uma das principais causas de evasão escolar, impactando psicologicamente os jovens e aumentando suas chances de abandonar os estudos.
9	SIGNIFICADO	Jovens que não se identificam com a escola a veem como uma perda de tempo e preferem dedicar-se a outras atividades. Ressignificar o currículo do Ensino Médio é uma maneira de atender às necessidades dos jovens e da sociedade.
10	FLEXIBILIDADE	Jovens se engajam menos nas atividades escolares quando sentem que a escola não é dinâmica ou inovadora. Para promover o interesse de todos, a flexibilidade em todos os aspectos da educação, desde o currículo até os métodos de ensino e avaliação, é essencial. Quanto mais flexível a escola, mais fácil é adaptar-se aos interesses e motivações dos alunos.
11	QUALIDADE DA EDUCAÇÃO	Tão importante quanto convencer os jovens da importância do que a escola ensina é garantir que o conteúdo seja verdadeiramente relevante para eles. Melhorar a qualidade da educação e investir no desenvolvimento dos professores, como incentivar turmas menores e oferecer formação continuada, são essenciais.
12	CLIMA ESCOLAR	Para garantir o engajamento dos jovens na escola e reduzir as chances de abandono, é vital que se sintam seguros, respeitados e pertencentes. Quando percebem que a escola foi pensada para eles e que é um lugar deles, sua motivação aumenta.
13	PERCEPÇÃO DA IMPORTÂNCIA	A escola deve não apenas ensinar temas relevantes, mas também motivar os jovens, mostrando a utilidade do que estão aprendendo e apresentando a educação como um valor. É importante que a escola e os pais deixem claro para os jovens a importância de estar na escola, para evitar decisões precipitadas de abandono.
14	BAIXA RESILIÊNCIA EMOCIONAL	Desentendimentos, baixo desempenho, problemas pessoais ou depressão podem levar ao desinteresse na escola. Identificar e resolver rapidamente esses problemas, com o apoio da comunidade escolar, pode evitar repercussões de longo prazo. Políticas para promover o engajamento dos jovens devem incluir acompanhamento e aconselhamento para aqueles em risco de desengajamento.

Tabela 2 – Causas de abandono escolar no Brasil

Fonte: (BARBOSA, 2017)

Segundo (MARQUES, 2020), os fatores que contribuem para a evasão incluem falta de comprometimento do aluno com o curso, falta de suporte familiar, pouca participação em atividades acadêmicas, baixo desempenho escolar, condições precárias nas Instituições de Ensino Superior (IES) e desmotivação em relação à carreira. Apesar do aumento

significativo no número de vagas nas IES nos últimos anos, a taxa de formatura não acompanha essa tendência, o que pode ser atribuído à alta taxa de evasão.

Sintetizando o ponto de vista de alguns autores, pode-se dizer, também que algumas das principais causas de evasão no ensino superior no Brasil, são:

1. Dificuldades financeiras: O alto custo da educação superior no Brasil pode ser uma barreira significativa para muitos estudantes. De acordo com a Pesquisa Nacional por Amostra de Domicílios (PNAD), em 2019, mais da metade dos estudantes universitários brasileiros vinham de famílias com renda per capita de até 1,5 salário mínimo. A falta de recursos financeiros pode levar os estudantes a atrasar ou interromper seus estudos ou até mesmo abandonar completamente a universidade (CAMPOS, 2016; OLIVEIRA; NÓBREGA, 2021; SACCARO; FRANÇA; JACINTO, 2019; FILHO et al., 2007).

2. Desafios acadêmicos: As demandas acadêmicas da universidade, como a carga de trabalho e as expectativas do corpo docente, podem ser um desafio significativo para alguns estudantes. Estudantes que não estão preparados para a transição da escola para a universidade ou que enfrentam dificuldades acadêmicas podem ser mais propensos a abandonar os estudos (CAMPOS, 2016; SACCARO; FRANÇA; JACINTO, 2019; OLIVEIRA; NÓBREGA, 2021; FILHO et al., 2007).

3. Falta de apoio social e psicológico: A transição para a universidade pode ser um momento estressante e isolado para muitos estudantes, especialmente aqueles que são os primeiros em sua família a frequentar a universidade. A falta de apoio social e psicológico pode contribuir para a evasão (SACCARO; FRANÇA; JACINTO, 2019; OLIVEIRA; NÓBREGA, 2021).

4. Escolha da carreira: Muitos estudantes ingressam na universidade sem uma compreensão clara de suas preferências e habilidades, o que pode levar a uma escolha equivocada de carreira ou curso. Os estudantes que percebem que escolheram a carreira ou curso errado podem ser mais propensos a abandonar a universidade (CAMPOS, 2016; SACCARO; FRANÇA; JACINTO, 2019; FILHO et al., 2007).

5. Problemas pessoais: Problemas pessoais, como problemas de saúde, familiares ou emocionais, podem levar os estudantes a abandonar a universidade (SACCARO; FRANÇA; JACINTO, 2019; MARQUES, 2020).

Segundo (PRESTES; FIALHO, 2018), é desafiador identificar os principais fatores que contribuem para a evasão escolar, mas os autores sugerem que a natureza desse fenômeno está ligada a aspectos psicológicos e individuais. Os fatores financeiros podem estar relacionados a questões familiares, socioculturais e de trabalho, enquanto os fatores acadêmicos podem envolver a trajetória escolar, métodos de ensino e estado emocional dos alunos.

2.2 Aprendizado de Máquina

Aprendizado de Máquina é um campo derivado da Inteligência Artificial e da Ciência da Computação, que busca capacitar sistemas computacionais a aprender e raciocinar de forma autônoma, imitando o pensamento humano. Isso é realizado por meio do desenvolvimento de algoritmos que podem acessar dados e informações e aprender de forma autônoma (LEMOS, 2021).

Para (SOUZA, 2020), a ML é um método de análise de dados que automatiza a criação de modelos analíticos a partir de dados históricos para tomar decisões, identificar padrões e fazer previsões com pouca intervenção humana.

A ML é uma área da inteligência artificial que tem ganhado destaque nos últimos anos. Essa área utiliza algoritmos e modelos estatísticos para analisar dados e fazer previsões. Com o uso de técnicas de aprendizado de máquina, é possível identificar padrões em dados complexos e fazer previsões precisas em diferentes áreas, como finanças, saúde, marketing e educação (ALPAYDIN, 2020).

Segundo (ESCOVEDO; KOSHIYAMA, 2020), a ML não é uma ciência nova, mas está ganhando destaque como uma ferramenta para analisar a crescente quantidade de dados em diversas áreas. Seus algoritmos aprendem com dados anteriores para identificar padrões, produzir decisões e resultados confiáveis e reproduzíveis. A natureza iterativa do aprendizado de máquina é crucial, pois permite que os modelos se adaptem de forma autônoma quando expostos a novos dados.

De acordo com (SOUZA, 2016), a ML utiliza modelos projetados para prever com precisão em novos conjuntos de dados, ou seja, mesmo sendo construídos com uma amostra específica, espera-se que tenham bom desempenho ao fazer previsões com novos dados inseridos.

(MITCHELL, 1997) explica que aprendizado de máquina envolve a exploração de um amplo conjunto de hipóteses para determinar aquela que melhor se ajusta aos exemplos de treinamento disponíveis, bem como a outras restrições ou conhecimentos prévios. Nesse sentido, o aprendizado de máquina busca responder à questão de como desenvolver algoritmos que aprimorem seu desempenho em determinada tarefa por meio da experiência.

Ainda segundo o autor, os algoritmos de ML têm demonstrado grande utilidade em uma ampla variedade de aplicações práticas, tais como: mineração de dados em grandes bancos de dados, identificação automática de regularidades implícitas, em domínios pouco compreendidos, nos quais os seres humanos podem não possuir o conhecimento necessário para desenvolver algoritmos eficazes e em áreas que exigem que o algoritmo se adapte dinamicamente às mudanças nas condições.

(SILVA; ZHAO, 2016), também explicam que a ML envolve o estudo e desenvolvimento de algoritmos que capacitam os computadores a aprender sem programação explícita. Essas técnicas podem ser aplicadas em diversas áreas, exigindo a tradução do problema para o domínio do Aprendizado de Máquina, que geralmente requer um conjunto de características como entrada e gera um critério de agrupamento ou classificação como saída.

2.2.1 Tipos de Aprendizado de Máquina

De acordo com Souza (2020), existem três categorias de aprendizado de máquina:

- Aprendizagem não supervisionada
- Aprendizagem supervisionada
- Aprendizagem semi-supervisionada.

Na aprendizagem não supervisionada, o algoritmo agrupa os dados com base em características semelhantes, sem a necessidade da variável a ser prevista. Na aprendizagem supervisionada, o algoritmo procura associações entre as variáveis preditoras e a variável a ser prevista em um conjunto de dados. Já na aprendizagem semi-supervisionada, é feita uma combinação da aprendizagem não supervisionada com a supervisionada, sendo mais utilizada em conjuntos de dados de grande volume (SOUZA, 2020; ENAP, 2020).

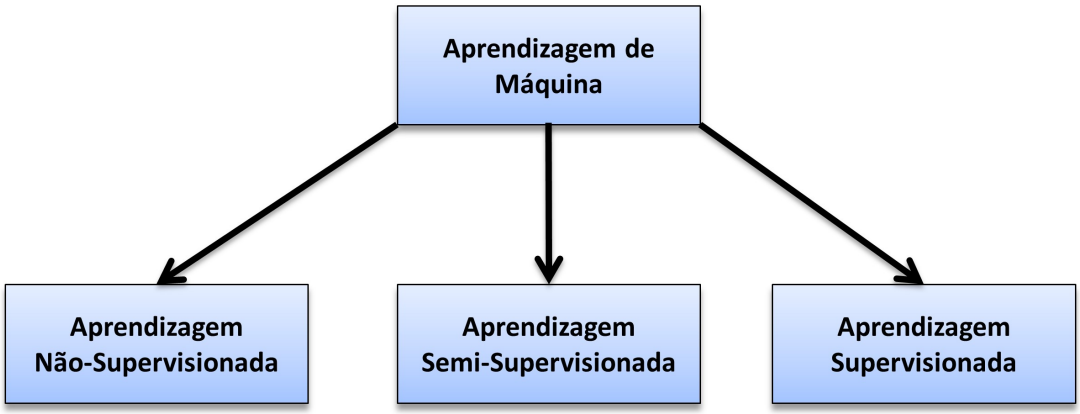

Figura 6 – Categorias de Aprendizagem de Máquina

Fonte: (SOUZA, 2020)

Neste trabalho serão utilizados algoritmos de aprendizagem supervisionada, uma vez que, será realizada uma modelagem preditiva controlada para a evasão escolar.

2.3 Técnica de Validação Cruzada K-fold

O K-Fold Validação Cruzada é uma técnica muito utilizada para classificação. Neste modelo o conjunto de dados é dividido em K partes sem sobreposição. O ciclo de

treinamento e validação é repetido K vezes, com cada uma das K partes sendo deixada de fora do treinamento uma vez, enquanto as partes restantes (K - 1) são usadas para validar o sistema. A parte retida é usada como conjunto de validação, e a média dos resultados dos K ciclos de avaliação é calculada para obter a estimativa final. A 7 ilustra o procedimento para o exemplo de K = 4 (RODRIGUEZ; PEREZ; LOZANO, 2009).

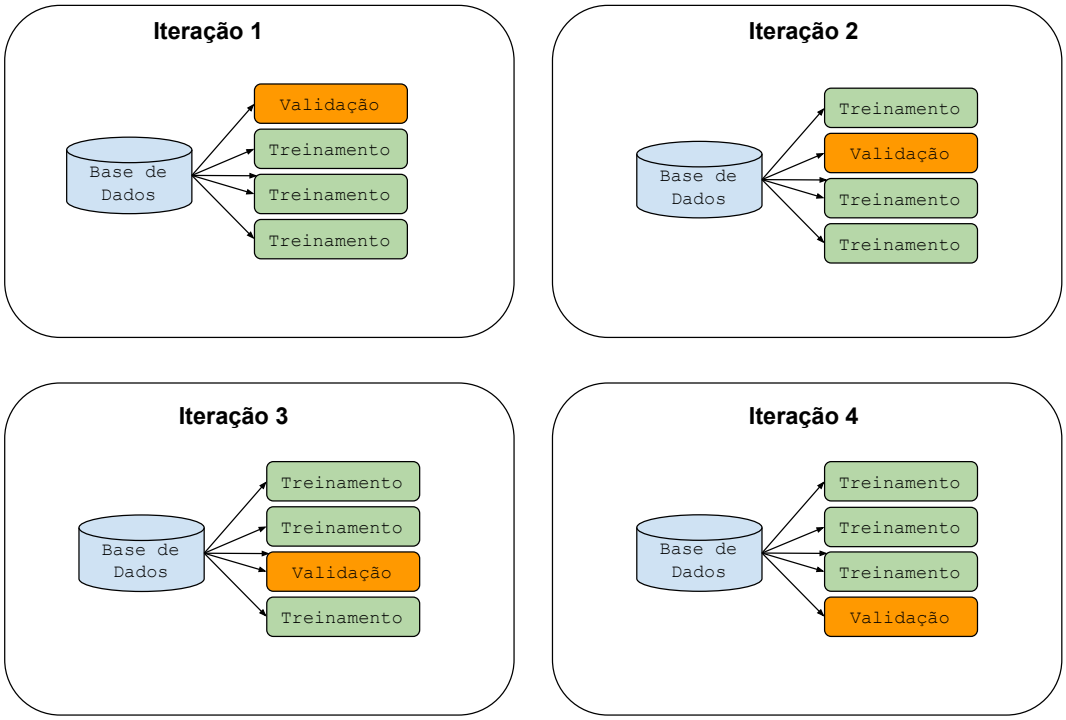

Figura 7 – Validação Cruzada para 4 folds.

Fonte: Elaborado pelo Autor

2.4 Classificadores de Aprendizado de Máquina

Existem várias técnicas de aprendizado de máquina, entretanto, nesse trabalho serão abordadas as tecnicas de aprendizagem supervisionada, as quais estão relacionadas aos modelos preditivos e as suas tarefas mais comuns são a classificação, que analisa dados de entrada, rotula e prevê uma categoria, como a aprovação ou reprovação de um aluno, e a regressão, que também analisa dados de entrada, mas determina um valor numérico (contínuo ou discreto) como saída (SOUZA, 2020).

Os modelos preditivos utilizam funções de aprendizado supervisionado para estimar valores desconhecidos de variáveis dependentes com base nas características das variáveis

independentes relacionadas. Eles têm o propósito específico de prever valores desconhecidos com base em valores conhecidos de outras variáveis. Essa previsão pode ser conceituada como um mapeamento de aprendizado a partir de um conjunto de entrada, como um vetor de medições, para uma saída, como um valor escalar (HAN; KAMBER; PEI, 2012).

De acordo com (SOUZA, 2020):

- Classificação: Trata-se de uma subcategoria de aprendizagem supervisionada que visa analisar a entrada e atribuir um rótulo a ela. Geralmente é utilizado quando as previsões são de natureza distinta, como um simples 0 ou 1. Seu resultado é um valor discreto, como determinar se um aluno evadiu ou não.

- Regressão: É uma outra subcategoria de aprendizagem supervisionada que busca modelar a relação entre os dados rotulados para determinar como o rótulo será alterado à medida que os valores dos recursos variam. É empregado quando o valor previsto não se limita a um simples 0 ou 1.

Segundo o mesmo autor, a diferença principal entre esses dois modelos preditivos é que a classificação prevê rótulos categóricos (discretos, não ordenados) para os dados, enquanto a regressão estabelece modelos de funções com valores contínuos.

2.4.1 Regressão Logística

A Regressão Logística é uma técnica crucial em aprendizado de máquina, frequentemente empregada para problemas de classificação. Conforme discutido por (JR; LEMESHOW; STURDIVANT, 2013), ela é uma extensão da regressão linear, adaptada para modelar a probabilidade de uma variável dependente categórica estar associada a um conjunto de variáveis independentes. Este método é fundamental para compreender e prever relações em dados onde a variável de resposta é discreta, sendo amplamente utilizada em diversas áreas, desde ciências médicas até finanças.

A regressão logística é utilizada para determinar a probabilidade de uma entrada pertencer a um determinado grupo, com aplicabilidade em problemas de classificação. Trata-se de algoritmos utilizados para prever variáveis com valores discretos, binários, como sim/não, verdadeiro/falso, a partir de um conjunto de variáveis independentes (ESCOVEDO; KOSHIYAMA, 2020; DELEN, 2011).

Em uma perspectiva mais abrangente de técnicas de aprendizado de máquina, a Regressão Logística é frequentemente incluída como um dos métodos fundamentais de aprendizagem supervisionada. De acordo com (HASTIE et al., 2009), a aprendizagem supervisionada, no qual a Regressão Logística se enquadra, é crucial para ensinar modelos a fazer previsões a partir de dados rotulados, estabelecendo relações entre inputs e outputs conhecidos.

Segundo (ALGHAMDI et al., 2017), a regressão logística é um classificador estatístico linear que estima a probabilidade de prever a classe rotulada em um tipo categórico usando vários atributos. Esse modelo de classificação mede a relação entre os atributos e a classe rotulada.

2.4.2 Árvores de Decisão (Decision Tree – DT)

A árvore de decisão (AD), do inglês Decision Tree (DT), é um modelo preditivo fundamental inspirado na forma humana de tomar decisões. Pode ser aplicada em problemas de classificação ou regressão (ESCOVEDO; KOSHIYAMA, 2020).

O método da árvore de decisão utiliza uma estrutura em forma de árvore para representar diferentes caminhos de decisão e seus resultados. Este modelo é de fácil interpretação e pode lidar com atributos numéricos e categóricos, além de classificar dados com atributos faltantes. Em problemas de regressão, a previsão para uma observação é a média ou moda das observações de treinamento da região correspondente. As regras de divisão dos segmentos são resumidas em árvores (GRUS, 2016).

De acordo com (SOUZA, 2020), um algoritmo de árvore de decisão é uma ferramenta preditiva útil para resolver problemas de regressão e classificação em várias áreas. Geralmente, um algoritmo de árvore de decisão divide uma base de dados em diferentes condições. Este método é prático e amplamente utilizado na aprendizagem supervisionada, onde o percurso da raiz até a folha representa uma regra de classificação.

Uma árvore de decisão, conforme (RUSSELL; NORVIG, 2010), é uma função que recebe um conjunto de valores como entrada e retorna uma única previsão da classe desejada. Essa previsão é determinada por uma série de testes. Cada nó interno representa um teste em um atributo de entrada, e os ramos que saem de cada nó são os possíveis valores desse atributo. Um nó sem ramos é chamado de folha e representa o valor retornado pela função.

Segundo (MITCHELL, 1997), as árvores de decisão classificam as instâncias, partindo da raiz até um nó folha que fornece a classificação final. Cada nó realiza um teste em um atributo da instância, resultando em várias divisões até chegar à folha, que contém a resposta final. Diversos algoritmos de árvore de decisão, como CHAID, CTree, C4.5, CART e Hoeffding Tree, compartilham semelhanças e são fundamentados em uma abordagem de divisão e conquista, com um processo recursivo. A distinção principal entre eles reside na seleção das variáveis e nos critérios de particionamento e de interrupção para o crescimento da árvore (ESCOVEDO; KOSHIYAMA, 2020).

A Figura 8 ilustra uma árvore de decisão.

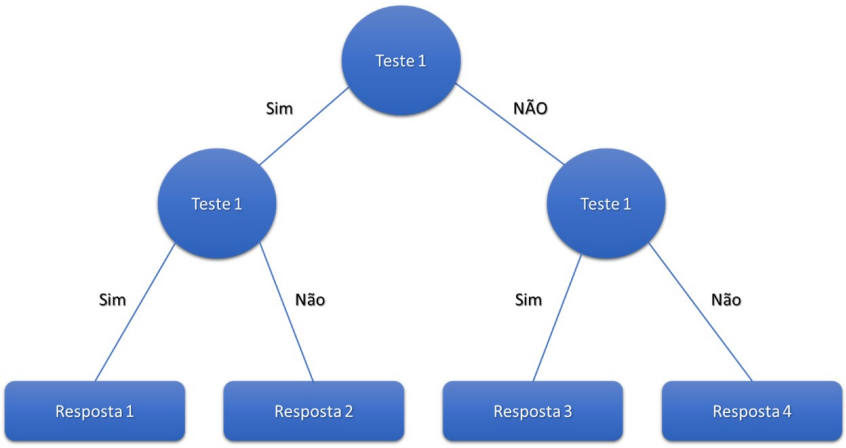

Figura 8 – Categorias de Aprendizagem de Máquina

Fonte: (ROLIM; CORDEIRO; FERREIRA, 2014)

2.4.3 Random Forest (Floresta Aleatória)

Random Forest é um algoritmo de Aprendizado de Máquina que utiliza várias árvores de decisão para produzir resultados mais precisos e robustos.

Conforme (SOUZA, 2020), um algoritmo de Random Forest, representa uma evolução do algoritmo de árvore de decisão. Ele envolve a construção de múltiplas árvores de decisão e, em seguida, combina suas saídas para aprimorar a capacidade de generalização do algoritmo.

De acordo com (RESENDE; DRUMMOND, 2018), Random Forest é um modelo composto por um grande número de árvores de decisão que atuam em conjunto. No caso da classificação, a previsão é baseada na maioria dos votos dos valores previstos em cada árvore de decisão.

Segundo (SILVA, 2022), Random Forest utiliza duas técnicas principais: bagging (ou bootstrap aggregating) e árvore de decisão. Esse algoritmo é comumente empregado em problemas de classificação e regressão.

O algoritmo de Random Forest consiste em várias árvores de decisão independentes que contribuem para o modelo final. Trata-se de um método de conjunto que combina diferentes abordagens para criar um modelo mais complexo e robusto, embora demande mais tempo computacional e ofereça resultados superiores. Nele, são criadas amostras aleatórias do banco de dados usando o método de reamostragem bootstrap, selecionando amostras com reposição dos elementos. Com base nesse princípio, uma árvore de decisão é modelada para cada amostra criada (TEODORO; KAPPEL, 2020).

Após o treinamento das árvores, elas recebem pesos com base em seu desempenho de previsão, e a classe resultante é calculada combinando todos os resultados multiplicados pelos pesos. A classe com a maior probabilidade ponderada é então determinada como

o valor de saída para o algoritmo. Esse método aumenta a complexidade e melhora a precisão das previsões (GÉRON, 2019).

A Figura 9 ilustra um modelo de Floresta Aleatória.

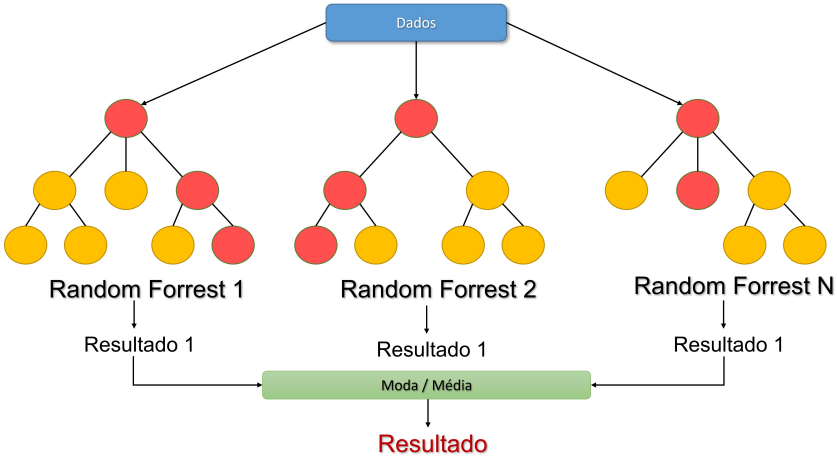

Figura 9 – Modelo de Random Forest

Fonte: Elaborado pelo próprio autor

2.4.4 Support Vector Machines (SVM)

Support Vector Machine (SVM), ou Máquina de Vetores de Suporte (MVS), é um algoritmo de classificação que busca encontrar a melhor fronteira para separar as classes, utilizando os vetores de suporte, ou seja, os exemplos mais próximos à borda de separação (PINHEIRO; SILVA; SOUZA, 2018).

Essa técnica foi idealizada por (CORTES; VAPNIK, 1995), e se baseia na ideia de separar os dados entre classes por meio de uma reta, plano ou hiperplano, dependendo do número de dimensões.

Segundo (WU et al., 2008), a SVM se destaca como um dos algoritmos mais robustos e precisos, apresentando vantagens como um sólido embasamento teórico, a necessidade de poucos exemplos durante o treinamento e insensibilidade ao número de atributos.

A SVM classifica um vetor de entrada em classes de saída conhecidas, buscando o hiperplano ideal que maximiza a separação entre as classes. Quando lidando com dados não linearmente separáveis, o método Kernel é utilizado para transformar o espaço de entrada em um espaço de recursos de alta dimensão, onde um hiperplano ideal linearmente separável pode ser construído (CHANG; LIN, 2011).

Apesar da aplicação dos modelos de SVM ser geralmente lenta, este é considerado um dos algoritmos mais eficientes, demanda poucos ajustes, tende a oferecer alta precisão

e é capaz de modelar fronteiras de decisão complexas e não lineares, de acordo com (ESCOVEDO; KOSHIYAMA, 2020).

Em resumo, a função de uma SVM é encontrar uma linha de separação, frequentemente chamada de hiperplano, entre os dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos de cada classe, como mostrado na figura 9. Essa distância entre o hiperplano e o primeiro ponto de cada classe costuma ser chamada de margem. A SVM coloca em primeiro lugar a classificação das classes, definindo assim cada ponto pertencente a cada uma das classes, e em seguida maximiza a margem. Ou seja, ela primeiro classifica as classes corretamente e depois em função dessa restrição define a distância entre as margens (MITCHELL, 1997). Como pode ser vista na figura 10.

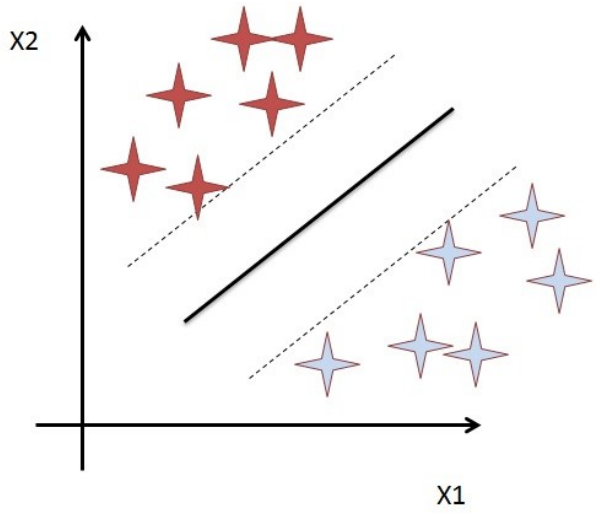

Figura 10 – Representação gráfica de SVM

Fonte: (MITCHELL, 1997)

2.4.5 Naïve Bayes

O algoritmo de classificação Naïve Bayes é um algoritmo que se fundamenta nas descobertas de Thomas Bayes para fazer previsões em aprendizado de máquina.

O termo "naive"(ingênuo) refere-se à maneira como o algoritmo analisa as características de um conjunto de dados, presumindo que, se o valor de um atributo afeta a distribuição de classes no conjunto, esse efeito é independente dos valores assumidos por outros atributos e de seus efeitos na mesma distribuição de classes (BIASI, 2019).

O classificador Naïve Bayes, juntamente com a árvore de decisão, é amplamente empregado na classificação de dados, pois demonstra um bom desempenho em problemas de classificação, independentemente se os dados são categóricos ou numéricos. Este classificador estático se baseia no Teorema de Bayes (BELLHOUSE, 2004).

Segundo (JURAFSKY; MARTIN, 2018), o algoritmo Naïve Bayes, foi desenvolvido em 1961 para tarefas de classificação, e é amplamente conhecido por seu uso na classificação de e-mails como mensagens normais ou de spam. Este algoritmo calcula probabilidades condicionais do tipo P(saída|entrada), ou seja, determina a probabilidade de uma determinada saída ocorrer dado uma entrada específica.

De acordo com (AMARAL, 2016), durante a etapa de treinamento, é criada uma tabela de valores onde é atribuído um peso a cada atributo em cada uma das classes de classificação. Ao submeter uma nova instância para classificação, o modelo irá somar os pesos de cada atributo em cada uma das classes, e a classe que obtiver a maior soma será a classe atribuída ao novo item.

O algoritmo Naïve Bayes é uma excelente escolha para conjuntos de dados extensos, pois oferece rapidez de execução em comparação com outros algoritmos de classificação. Esse tipo de classificador assume que a presença de uma característica específica em uma classe não está relacionada à presença de qualquer outra característica (SÁ et al., 2018).

Para (FACELI et al., 2021), o classificador Naïve Bayes geralmente demonstra alto desempenho devido ao fato de considerar cada atributo de forma independente, permitindo seu uso com informações incompletas e imprecisas.

2.4.6 K-Nearest Neighbors (k-NN)

O algoritmo KNN pertence à família de algoritmos de Instance-based Learning (IBL). Esses algoritmos armazenam todas as instâncias de treinamento e, ao classificar uma nova instância, recuperam um conjunto de instâncias semelhantes do conjunto de treinamento para auxiliar na classificação (FARIA, 2016).

A técnica kNN, do inglês k-Nearest Neighbours, que significa k Vizinhos mais próximos, é um dos algoritmos utilizados para aprendizagem supervisionada, tanto para problemas de classificação quanto de regressão. É um algoritmo simples de entender e não paramétrico, pois não faz suposições sobre a distribuição dos dados (ESCOVEDO; KOSHIYAMA, 2020).

Segundo (OLIVEIRA, 2016), o kNN é composto por três elementos principais: um conjunto de exemplos rotulados (por exemplo, um conjunto de registros armazenados), uma métrica de distância e o valor de k (o número de vizinhos mais próximos).

De acordo com (WU et al., 2008), o kNN é um método iterativo simples que identifica um grupo de k objetos do conjunto de treinamento que mais se assemelham a um novo dado em teste e classifica essa nova instância com base na classificação mais frequente nessa vizinhança.

(MEDEIROS et al., 2019) apresenta um exemplo para demonstrar como o algoritmo

kNN funciona, conforme figura 11, a seguir.

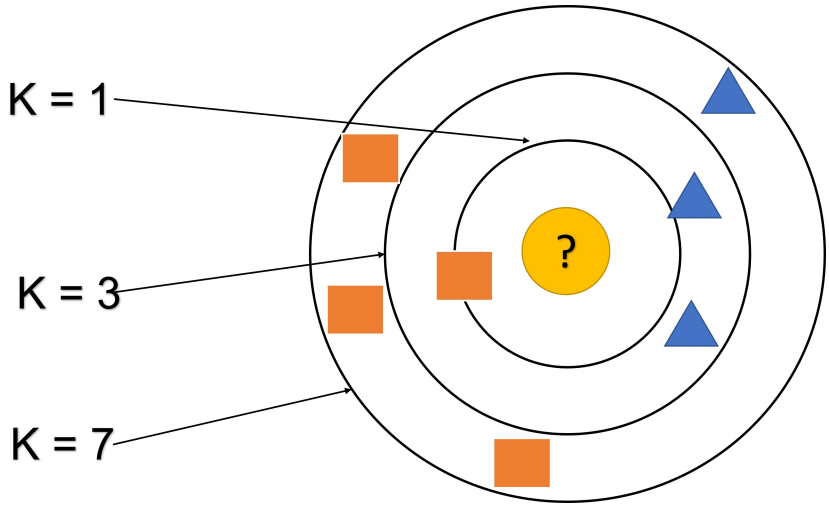

Figura 11 – Exemplo de Funcionamento do kNN

Fonte: (MEDEIROS et al., 2019)

2.4.7 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA), ou do inglês, Artificial Neural Networks (ANN), são algoritmos inspirados no modelo biológico do cérebro humano. Assim como no cérebro, as ANNs consistem em um conjunto de unidades, ou neurônios, interconectados por arestas chamadas sinapses. Essas sinapses têm pesos que são ajustados para minimizar o erro na saída da rede, o que constitui o processo de aprendizado (SOUZA, 2020). (MITCHELL, 1997) afirma que as RNAs são construídas a partir de um conjunto densamente interconectado de neurônios simples, onde cada neurônio leva um número de entradas de valor real (possivelmente as saídas de outros neurônios e produz uma única saída de valor real (que pode se tornar a entrada para muitos outros neurônios).

As Redes Neurais Artificiais (RNAs) podem ser aplicadas em problemas de classificação e regressão. De acordo com (GÉRON, 2019), o uso de RNAs supera muitas outras técnicas de aprendizado de máquina ao lidar com problemas grandes e complexos, podendo ser treinadas em um tempo razoável. Apesar de as redes neurais artificiais se inspirarem nos neurônios biológicos, é importante lembrar que suas semelhanças são mínimas. A intenção original era imitar a realidade biológica (FERREIRA, 2020).

O neurônio artificial é uma estrutura lógico-matemática que procura simular a forma, o comportamento e as funções de um neurônio biológico. Assim sendo, os dendritos foram substituídos por entradas, cujas ligações com o corpo celular artificial são realizadas através de elementos chamados de peso (simulando as sinapses). Os estímulos captados pelas entradas são processados pela função de soma, e o limiar de disparo do neurônio biológico foi substituído pela função de transferência.

2.5 Métricas de Avaliação

Nesta seção serão demonstradas as principais métricas de avaliação utilizadas para medir o desempenho de modelos e sistemas em diversas áreas da aprendizagem de máquina.

2.5.1 Matriz de confusão

A matriz de confusão é um teste para comparar qual classificação o algoritmo previu com a classificação real da instância, fornece um detalhamento do desempenho do modelo. Indica nas colunas as classes reais e nas linhas são identificadas as classes preditas (COSTA, 2021). Os termos utilizados na composição de uma matriz de confusão são:

- Verdadeiro Positivo (VP): número de exemplos positivos classificados corretamente;
- Falso Negativo (FN): número de exemplos negativos classificados incorretamente;
- Falso Positivo (FP): número de exemplos positivos classificados incorretamente;
- Verdadeiro Negativo (VN): número de exemplos negativos classificados corretamente.

2.5.2 Acurácia

A métrica acurácia é reconhecida como uma métrica simples, porem uma das mais importantes. Ela simplesmente avalia a porcentagem de previsões corretas, ou seja, pode ser calculada pela divisão da quantidade de previsões corretas pelo total de previsões realizadas. A Equação 2.1 mostra como calcular a acurácia:

$$acurácia = \frac{VP + VN}{VP + FN + VN + FP} \tag{2.1}$$

2.5.3 Precisão

A precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos, conforme equação 2.2:

$$precisão = \frac{VP}{VP + FP} \tag{2.2}$$

2.5.4 Sensibilidade

Esta métrica avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. Ela pode ser obtida pela equação **??**:

$$sensibilidade = \frac{VP}{VP + FN} \tag{2.3}$$

2.5.5 F1 Weighted (F1 Ponderado)

É uma medida que calcula a média ponderada do F1-score de cada classe, levando em conta o desequilíbrio entre elas. O F1-score é uma métrica que combina precisão e revocação em uma única medida. Ela pode ser obtida com base na equação **??**:

$$f1 = 2 * \frac{precisão * sensibilidade}{precisão + sensibilidade} \tag{2.4}$$

3 Trabalhos Relacionados

Neste capítulo são apresentados alguns trabalhos relacionados a esta pesquisa, que buscaram realizar a previsão de evasão universitária através de Aprendizagem de Máquina.

O modelo de (TINTO, 1982; TINTO, 1975) foi amplamente utilizado como referencial teórico na maioria dos trabalhos realizados devido à sua consideração dos fatores de sucesso acadêmico. Tinto descreve o processo de desgaste do estudante como uma interação socio-psicológica entre as características do aluno ao ingressar na universidade e sua experiência na instituição. Esta interação resulta em um nível de integração do aluno no novo ambiente, o qual está diretamente relacionado ao comprometimento com a instituição educacional e a conclusão dos estudos.

Estudos posteriores operacionalizaram esse modelo buscando uma modelagem preditiva controlada para a evasão no ensino superior.

(KARAMOUZIS; VRETTOS, 2008), conduziram uma pesquisa onde o foco foi a previsão da taxa de aprovação dos estudantes nos primeiros dois anos dos cursos de graduação. Eles empregaram uma rede neural Multilayer Perceptron (MLP) de três camadas, combinando variáveis demográficas e acadêmicas para criar um conjunto de dados com diversas variáveis. O estudo revelou uma taxa média de precisão de 72%.

(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009) conduziram um estudo experimental para desenvolver um modelo preditivo de evasão no curso de Engenharia Elétrica da Universidade de Tecnologia de Eindhoven, na Holanda. A taxa de evasão após o primeiro ano era de aproximadamente 40%. O estudo definiu o aluno evasor como aquele que, após 3 anos do ingresso, não havia concluído com êxito as matérias do primeiro ano. Os autores analisaram dados anteriores ao ingresso na universidade e dados de desempenho durante a universidade. Foram treinados diferentes algoritmos, incluindo regressão logística, OneR, algoritmos de árvores de decisão (CART, C4.5), classificador bayesiano (Naïve Bayes), algoritmo baseado em aprendizagem de regras e florestas aleatórias (Random Forest). E como resultado, os algoritmos que alcançaram maior acurácia foram os de árvores de decisão (CART, C4.5 e Florestas Aleatórias) e o algoritmo de regressão logística, com valores em torno de 80%. Os resultados indicaram que um modelo de Árvore de Decisão alcançou uma precisão de 68% ao analisar apenas os dados pré-universitários. Ao considerar o conjunto completo de dados, o método obteve uma precisão entre 75% e 80% na identificação da evasão.

(JADRIĆ; GARAČA; ČUKUŠIĆ, 2010) realizaram um estudo com estudantes iniciantes no curso de Economia, limitando-se aos primeiros dois anos. O objetivo era testar e comparar os modelos de classificação Árvore de Decisão, Regressão Logística

e Redes Neurais, utilizando a metodologia SEMMA (Sampling, Exploring, Modifying, Modelling and Assessment). O conjunto de dados foi construído com variáveis relacionadas à inscrição do candidato e atributos do processo de estudos. Os resultados revelaram que o modelo de rede neural apresentou melhor desempenho em comparação aos outros, identificando que 36% dos estudantes poderiam evadir. Além disso, a pesquisa identificou e diferenciou as causas da evasão, delineando o perfil típico do estudante propenso a desistir da faculdade.

(DELEN, 2011) analisou dados financeiros, acadêmicos e demográficos de mais de 25 mil estudantes em uma universidade pública nos Estados Unidos. Dos 25 mil alunos, 19 mil continuaram após o primeiro ano, resultando em uma taxa de evasão de aproximadamente 21%. Utilizando o histórico do ensino médio e do primeiro ano na universidade, o autor treinou classificadores e obteve uma precisão de 81% em redes neurais, 78% com árvore de decisão e 74% com regressão logística. O estudo também identificou que o desempenho acadêmico do aluno, tanto no presente quanto no passado, é um dos principais fatores relacionados à evasão.

(BALANIUK et al., 2011) propuseram a utilização de três algoritmos de classificação para categorizar os alunos com evasão, com base nos dados de mais de 11 mil estudantes de três cursos de uma instituição de ensino superior em Brasília. Os modelos foram treinados com atributos que incluem informações socioeconômicas e acadêmicas dos alunos. Como resultado, foi constatado que é viável identificar estudantes com alto risco de evasão com uma precisão de até 80.6

(MUSTAFA; CHOWDHURY; KAMAL, 2012) conduziram um estudo que teve como objetivo prever a evasão de estudantes por meio de um modelo dinâmico. Foram utilizados os modelos de Árvores de Classificação e Regressão (CART) e CHAID. O conjunto de dados incluiu variáveis como sexo, situação financeira, período do curso em que houve evasão e presença de deficiência. Os resultados indicaram que a utilização apenas dos dados de inscrição dos candidatos não é ideal para identificar a evasão, com taxas de precisão de 38,1% para a Árvore CHAID e 28,57% para a Árvore CART. Os autores recomendam incluir outros fatores, como idade, etnia, situação de trabalho, ambiente de estudo e tipo de educação, para melhorar o desempenho na identificação da evasão.

(MANHÃES et al., 2011), exploraram exclusivamente os atributos extraídos dos registros acadêmicos dos alunos. Eles utilizaram cinco algoritmos de classificação e dados de seis cursos da Universidade Federal do Rio de Janeiro. Essa abordagem resultou em uma precisão de pelo menos 87% para cada curso.

(AULCK et al., 2016), analisaram, em sua pesquisa, as características-chave que indicam evasão. O foco principal foi prever a evasão do aluno com base em informações demográficas e registros escolares. Os autores utilizaram a base de dados da Universidade de Washington, contendo informações de estudantes de graduação entre 1998 e 2006,

totalizando cerca de 69 mil registros. Eles empregaram três algoritmos de classificação (Regressão logística regularizada, KNN e Random Forest), obtendo desempenhos distintos para cada algoritmo.

(MARIA; DAMIANI; PEREIRA, 2016), realizaram um estudo onde foi empregado o uso de redes bayesianas para prever o percentual de chances de evasão de estudantes. Eles coletaram características dos alunos do sistema utilizado pelo SENAI/SC. A validação dos resultados revelou uma taxa de acerto de 85,6%, demonstrando o bom desempenho da rede bayesiana modelada para o sistema desenvolvido.

No estudo de (ASSIS, 2018), foi utilizada a mineração de dados para identificar o perfil dos alunos evasores, relacionando dados do curso, área de estudo e instituição. O autor combinou informações do censo da educação superior e do ENEM para criar 5 modelos de algoritmos classificatórios: naïve bayes, redes neurais, regressão logística e dois algoritmos de árvores de decisão (CART e C5.0). O algoritmo CART se destacou em relação aos outros classificadores, alcançando uma sensibilidade de 84% na classificação dos alunos evadidos, enquanto nos demais testes os resultados não foram estatisticamente significativos ao comparar os algoritmos.

(PINHEIRO; SILVA; SOUZA, 2018) utilizaram, por meio das técnicas de Aprendizagem de Máquina, os algoritmos de Naïve Bayes (NB), Support Vector Machines (SVMs) e Árvores de Decisão (AD) para identificar precocemente quais alunos têm maior propensão à evasão. Os dados foram coletados do Sistema Acadêmico do Instituto Federal de Educação Ciência e Tecnologia do Maranhão e incluíram informações socioeconômicas e acadêmicas dos alunos.

(MELO et al., 2019) empregou a mineração de dados por meio de redes neurais artificiais para prever a possibilidade de abandono do ensino superior. Utilizando dados do Sistema Acadêmico da Universidade Federal do Triângulo Mineiro, a ferramenta desenvolvida foi capaz de identificar 63,8% dos alunos que abandonaram e prever, em média, com 36 dias de antecedência, os alunos propensos a evadir. Isso possibilitou que a instituição atuasse de forma preventiva.

(MOREIRA, 2020) analisou, durante o período de 2011 a 2019, a trajetória dos alunos do curso de ciência da computação da UFPR, utilizando duas técnicas de aprendizado de máquina: árvore de decisão e regressão logística, com o objetivo de prever a evasão. Com os dados do primeiro ano, foi possível prever a evasão com uma precisão de 74% e sensibilidade de 85% usando árvore de decisão, e com uma precisão de 75% e sensibilidade de 85% com regressão logística.

(TEODORO; KAPPEL, 2020) buscaram identificar os padrões característicos e os atributos mais determinantes dos alunos com maior potencial de desvinculação das instituições de ensino superior, aplicando cinco técnicas de aprendizado de máquina (Naïve

Bayes, K-Nearest Neighbors, Árvores de Decisão, Random Forest e Redes Neurais) nos dados do INEP. Onde, a Random Forest obteve o melhor resultado, com uma taxa de acerto de aproximadamente 80%, e forneceu dados que possibilitam a geração de um relatório sobre os atributos mais importantes para suas previsões.

(SOUZA, 2020) procurou apresentar, em seu estudo, uma ferramenta para combater à evasão usando as técnicas de aprendizado de máquina: árvores de decisão com acurácia de 96,4%, SVM com 76,4%, regressão logística com 77,7%, redes neurais com 94,4%, entre outras.

(LEMOS, 2021) utilizou os algoritmos de árvores de decisão, Random Forest e XGBoost, como ferramenta para prever a evasão escolar. Onde o algoritmo Random Forest obteve a melhor acurácia com 64%, seguida do algoritmo XGBoost.

(COSTA, 2021), utilizou em seu estudo a análise dos três primeiros semestres cursados para criar modelos de previsão do risco de evasão de alunos nos cursos de ciência da computação e engenharias na Universidade Federal de Pelotas (UFPel). Foram utilizados dados de 22 atributos extraídos do sistema acadêmico da universidade e aplicados em cinco algoritmos. Para o curso de ciência da computação, o algoritmo com maior precisão foi a regressão logística (90,16%), enquanto para os dados das doze engenharias, o modelo de floresta aleatória foi o mais preciso (83,4%).

(PRIMÃO et al., 2022) propôs um modelo usando algoritmos de aprendizado de máquina para prever a evasão escolar no Instituto Federal de Santa Catarina (IFSC). Para o desenvolvimento do modelo, foi utilizada a metodologia CRISP-DM, como modelo baseline, foi utilizado o algoritmo DecisionTreeClassifier e, para o desenvolvimento do modelo, os algoritmos XGBClassifier e MLPClassifier.

(BARBOSA, 2021) realizou um estudo com base nas informações do banco de dados do Sistema de Gestão Acadêmica da Universidade Federal do Paraná (SIGA/UFPR), comprovando que as técnicas de Aprendizado de Máquina podem ser satisfatoriamente empregadas na mineração de dados. Isso demonstra que a regressão logística, árvore de decisão, kNN, SVM e Random Forest são classificadores que viabilizam a identificação automática de alunos propensos à evasão. Essas informações podem auxiliar a gestão na tomada de decisões após o término do primeiro semestre, permitindo o desenvolvimento de estratégias eficazes de combate à evasão e promovendo uma mudança organizacional significativa.

Como observado, os estudos relacionados demonstram resultados positivos com algoritmos de classificação, e corroboram para a aplicação de técnicas de ML na prevenção de evasão escolar, objeto de estudo desse trabalho.

4 Metodologia

Neste capítulo, é apresentada a caracterização da pesquisa, juntamente com a descrição detalhada do procedimento adotado e de suas etapas.

A Figura 12 apresenta um esquema de funcionamento do modelo proposto neste trabalho (técnica de validação cruzada K-Fold 2.3).

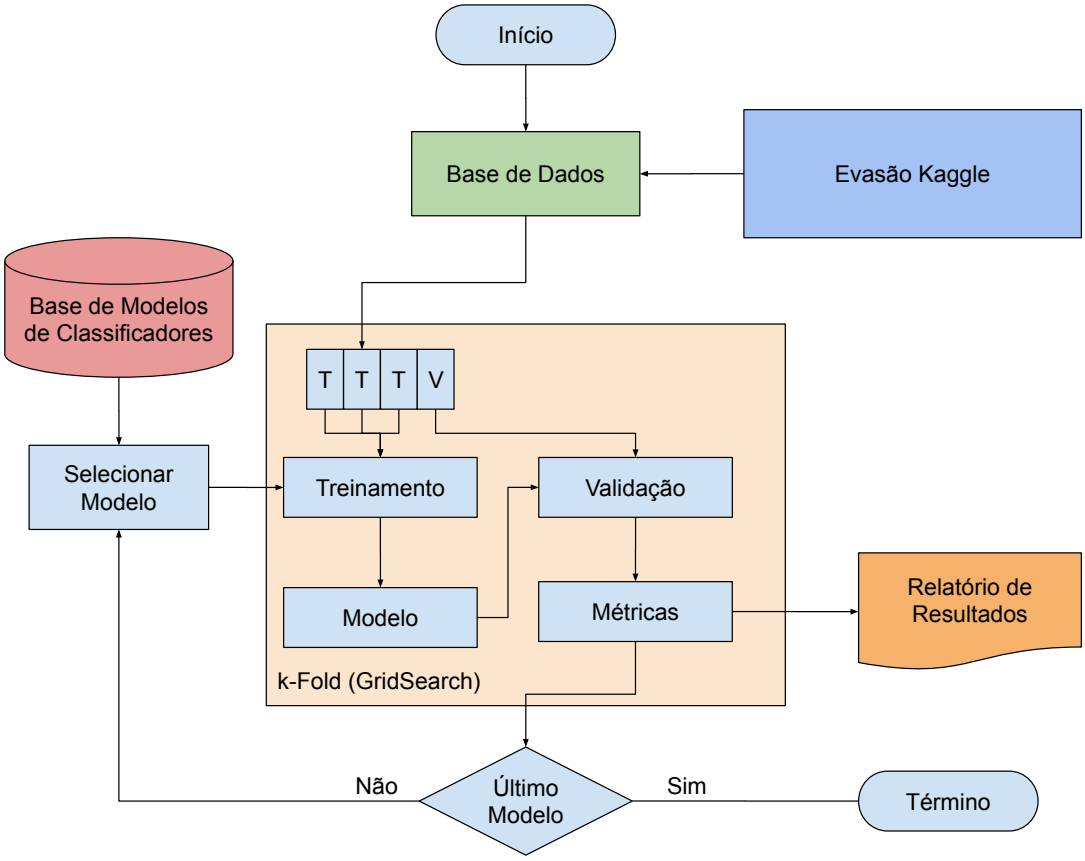

Figura 12 – Etapas do processo do modelo ML

Fonte: Elaborado pelo Autor

4.1 Caracterização da Pesquisa

(GIL et al., 2002) descreve pesquisa como um procedimento racional e sistemático com o propósito de oferecer respostas aos problemas propostos. De acordo com o autor, a pesquisa envolve diversas fases, desde a formulação do problema até a apresentação e discussão dos resultados.

Conforme (NASCIMENTO; SOUSA, 2016), a pesquisa pode ser distinguida pela sua natureza, pelos métodos (abordagem metodológica), pelos objetivos e pelos proce-

dimentos. Assim, o presente estudo consiste em uma pesquisa de natureza **aplicada**, dedicada à geração de conhecimento para solução de problemas específicos, com abordagem **qualitativa**, baseada na interpretação dos fenômenos observados e em seu significado, e com objetivo **explicativo**, devido à sua maior complexidade e ao emprego de método experimental de pesquisa. Quanto aos procedimentos, se caracteriza com uma **pesquisa experimental**.

Sendo assim, é preciso compreender claramente qual é o objetivo do projeto e como ele se relaciona com as necessidades do negócio ou da aplicação em questão. Isso envolve definir o que se deseja alcançar com o modelo de aprendizado de máquina e como o seu desempenho será avaliado. A partir daí, é possível avançar para as etapas seguintes do processo.

1. **Coleta de dados:** Nesta etapa, são reunidos os dados relevantes para o problema de aprendizado de máquina. Obtidos através de fontes diversas, como bancos de dados, arquivos, APIs entre outras fontes de informação. A qualidade e a quantidade dos dados coletados são fundamentais para o sucesso do modelo de aprendizado de máquina.

2. **Preparação dos Dados:** Após a coleta dos dados, estes são preparados para serem utilizados no modelo de aprendizado de máquina. Incluindo tarefas como limpeza dos dados (remoção de valores ausentes ou duplicados), normalização (garantindo que os dados estejam em uma escala consistente) e seleção de características relevantes para o problema em questão.

3. **Divisão dos dados:** Inicialmente, o conjunto de dados é dividido em k partes (ou folds) de tamanhos aproximadamente iguais. Para este modelo foi usado o parâmetro de 3 partes.

4. **Treinamento e validação:** O modelo é treinado k vezes. Em cada iteração, um dos folds é retido como conjunto de validação e os outros k-1 folds são usados como conjunto de treinamento.

5. **Avaliação:** Após as k iterações, são calculadas as métricas de desempenho do modelo (como acurácia, precisão, recall, etc.) em cada fold de validação.

6. **Média dos resultados:** Finalmente, os resultados obtidos em cada fold de validação são combinados (por exemplo, calculando a média) para obter uma estimativa mais robusta do desempenho do modelo.

7. **Implantação do Modelo:** A otimização de parâmetros busca constantemente aprimorar a qualidade e eficiência do modelo de aprendizado de máquina em uso,

identificando valores que impactam diretamente na precisão do modelo e o tempo de treinamento necessário.

8. **Análise dos Resultados:** É o momento em que as técnicas de aprendizado de máquina pode ser efetivamente utilizada para fornecer respostas às perguntas para as quais foi treinada.

4.2 Ferramentas e Técnicas

Durante a realização deste trabalho, foram empregadas algumas ferramentas e técnicas fundamentais.

Foi utilizada a linguagem de programação *Python* em conjunto com o ambiente do *Google Colab*, representado na figura 13, que proporcionou recursos computacionais na nuvem e facilitou o desenvolvimento do projeto.

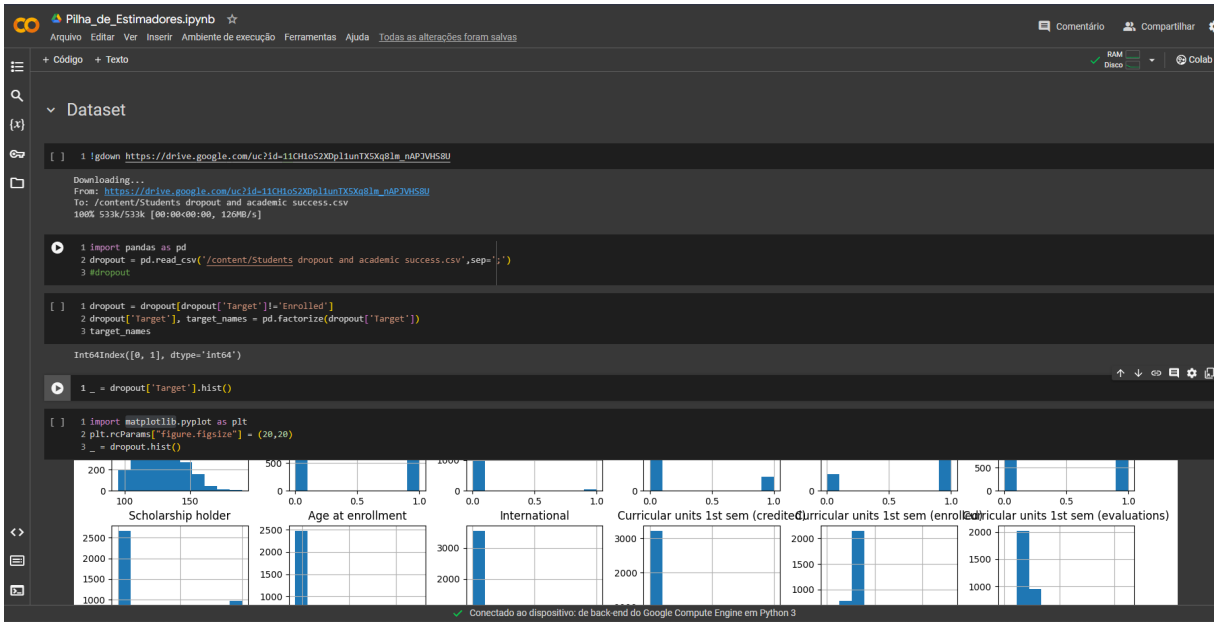

Figura 13 – Ambiente de Desenvolvimento do Google Colab

Fonte: Elaborado pelo próprio autor

A biblioteca Pandas, uma biblioteca para Ciência de Dados de código aberto (*open source*), construída sobre a linguagem *Python*, e que providencia uma abordagem rápida e flexível, com estruturas robustas para se trabalhar com dados relacionais (ou rotulados), de maneira simples e intuitiva, foi fundamental para a manipulação e análise eficiente dos dados.

O *Scikit-learn (sklearn)*, uma biblioteca da linguagem Python desenvolvida especificamente para aplicação prática de aprendizado de máquina, ofereceu uma ampla gama

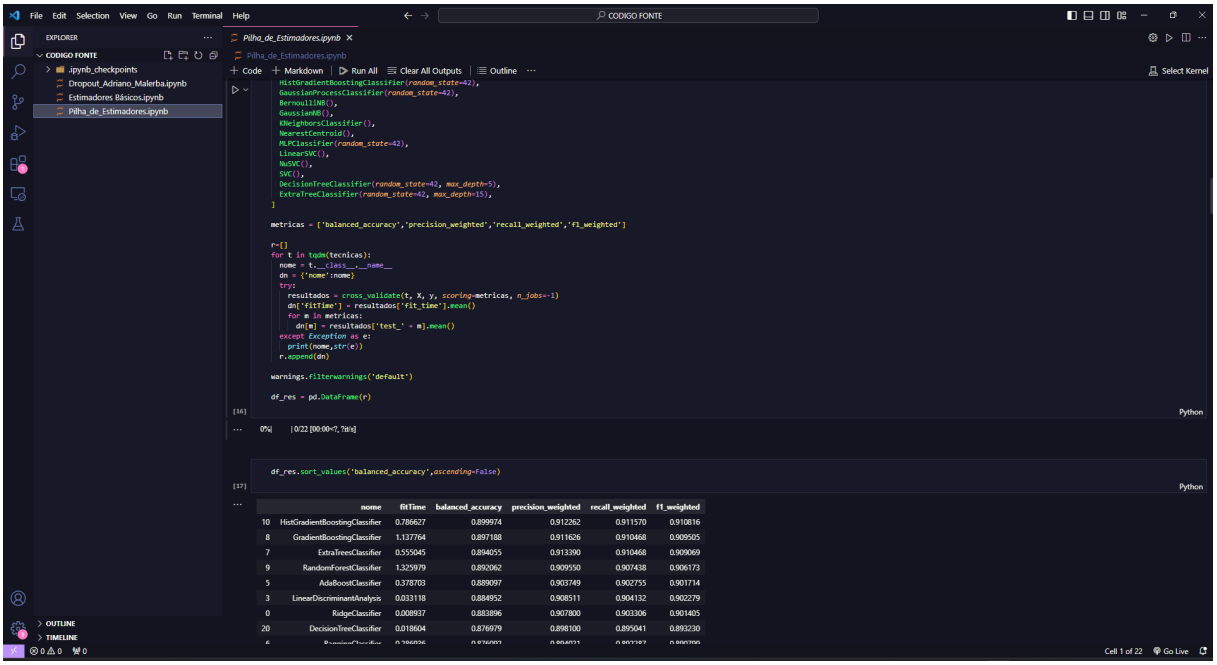

Figura 14 – Ambiente de Desenvolvimento do VSCode da Microsoft

Fonte: Elaborado pelo próprio autor

de algoritmos de aprendizado de máquina para a construção dos modelos preditivos.

A biblioteca *Matplotlib*, uma biblioteca de visualização de dados em *Python*, que oferece uma vasta gama de recursos para criar gráficos estáticos, animações, gráficos 3D e muito mais, foi utilizada para a visualização dos resultados e criação de gráficos.

O uso da linguagem SQL foi essencial para consultas e manipulação de dados no contexto de modelos relacionais.

O ambiente de desenvolvimento *Visual Studio Code* (VSCode), demonstrado na figura 14, proporcionou uma plataforma robusta para a escrita de código e gerenciamento do projeto.

Por fim, a integração com o Sistema de Gerenciamento de Banco de Dados PostgreSQL (SGBD PostgreSQL) permitiu o armazenamento e recuperação eficiente dos dados utilizados no projeto.

A combinação dessas ferramentas e técnicas foi essencial para a realização deste trabalho e contribuiu significativamente para a obtenção dos resultados alcançados.

4.3 Base de Dados Pública

A análise de dados é uma parte essencial do processo de tomada de decisão. Com o crescimento do volume de dados disponíveis, a necessidade de técnicas eficazes de análise

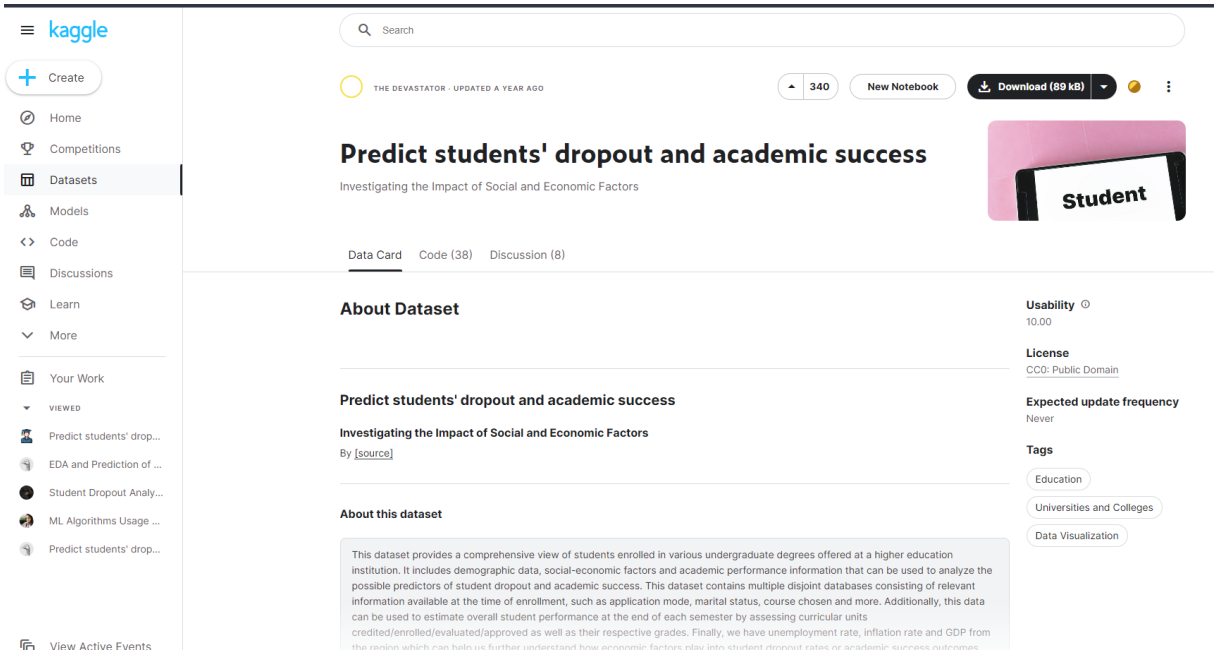

Figura 15 – Kaggle Plataforma de conjunto de dados

Fonte: (KAGGLE, 2021)

de dados tornou-se cada vez mais crucial. Neste contexto, a plataforma Kaggle tem desempenhado um papel significativo ao fornecer conjuntos de dados e competições que permitem aos cientistas de dados e analistas explorar e analisar informações de diversas fontes. Nesta dissertação, será discutido o processo de análise de dados realizada com base nos conjuntos de dados obtidos na plataforma Kaggle representada na figura 15.

O primeiro passo no processo de análise de dados foi a coleta dos conjuntos de dados relevantes da plataforma Kaggle. A escolha dos conjuntos de dados foi baseada nos objetivos da análise e nas questões que se pretendia responder.

O Kaggle foi escolhido por ser uma plataforma que permite aos usuários encontrar e publicar conjuntos de dados que podem ser usados para compreender e prever a evasão de estudantes e os resultados acadêmicos. Os dados incluem uma variedade de fatores demográficos, socioeconômicos e de desempenho acadêmico relacionados aos alunos matriculados em instituições de ensino superior.

Neste conjunto de dados, existem 4.424 registros de informações sobre alunos, cada um contendo 34 características, além da variável alvo que indica se o aluno evadiu ou não, além de outras informações relevantes disponíveis no momento da inscrição, como modo de inscrição, estado civil, curso escolhido e muito mais.

Com sua utilização, é possível investigar duas questões-chave: 1) Quais fatores preditivos específicos estão vinculados à evasão ou conclusão do aluno? 2) Como diferentes características interagem entre si?

Também é possível explorar se existem características demográficas (por exemplo, gênero, idade na matrícula etc.) ou condições de imersão (por exemplo, taxa de desemprego na região) associadas a taxas mais altas de sucesso do aluno, bem como entender quais implicações a condição financeira tem para os resultados educacionais.

Ao responder a essas perguntas, são gerados *insights* de pesquisa que podem fornecer informações críticas para os administradores na formulação de estratégias que promovam a conclusão bem-sucedida do curso entre alunos de origens diversas em suas instituições.

Para usar este conjunto de dados de forma eficaz, é importante estar familiarizado com todas as variáveis fornecidas, incluindo variáveis categóricas (qualitativas) como gênero ou modo de inscrição; variáveis numéricas como número de unidades curriculares no início dos semestres ou idade na matrícula; variáveis do tipo de medição de dados ordinais como estado civil; tendências estudadas ao longo do tempo, como taxa de inflação ou PIB; variáveis de medição de frequência como porcentagem de bolsistas; etc.

Coletar dados da base de dados Kaggle foi uma escolha estratégica devido à diversidade e qualidade dos conjuntos de dados disponíveis, proporcionando uma ampla gama de informações para análise. Além disso, a plataforma Kaggle oferece uma comunidade ativa e recursos valiosos, como competições e discussões, que ajudam a aprimorar a análise de dados e a desenvolver habilidades em ciência de dados. A coleta de dados do Kaggle foi realizada por meio do download direto dos conjuntos de dados relevantes disponíveis na plataforma, garantindo acesso a informações confiáveis e atualizadas para a análise.

4.3.1 Preparação dos Dados

Após a coleta, os dados foram submetidos a um processo de limpeza e pré-processamento. Isso incluiu a remoção de valores ausentes, tratamento de *outliers* e normalização, garantindo que os dados estivessem prontos para a análise.

Durante esta fase, o objetivo foi obter um entendimento mais aprofundado da fonte de dados, seus atributos e aspectos importantes, incluindo a detecção de anomalias, tipos de dados, quantidade de registros, entre outros.

4.3.2 Caracterização de Atributos

Nesta etapa, o objetivo foi otimizar os algoritmos de aprendizado de máquina por meio da transformação dos dados. Para isso, foram realizados testes com os 35 atributos selecionados na base de dados final. A tabela 3 apresenta a base de dados final utilizado nos modelos de aprendizado de máquina.

Campo	Descrição	Tipo de Variável
Estado Civil	O estado civil do aluno	Categórica
Modo de aplicação	O método de aplicação utilizado pelo aluno	Categórica
Ordem de inscrição	Ordem em que o aluno se inscreveu	Numérica
Curso	O curso feito pelo aluno	Categórica
Frequência diurna/noturna	Se o aluno frequenta as aulas durante o dia ou à noite	Categórica
Habilitação anterior	Qualificação obtida pelo aluno antes de ingressar no ensino superior.	Categórica
Nacionalidade	Nacionalidade do estudante	Categórica
Qualificação da mãe	Qualificação da mãe do aluno	Categórica
Qualificação do pai	Qualificação do pai do aluno	Categórica
Ocupação da mãe	Ocupação da mãe do aluno	Categórica
Ocupação do pai	Ocupação do pai do aluno	Categórica
Refugiado	Se o aluno é uma pessoa Refugiado	Categórica
Necessidades educacionais especiais	Se o aluno tem alguma necessidade educacional especia	Categórica
Devedor	Se o aluno é devedor	Categórica
Mensalidades em dia	Mensalidades do aluno estão em dia	Categórica
Gênero	O gênero do aluno	Categórica
Bolsista	Se o aluno é bolsista	Categórica
Idade na matrícula	A idade do aluno no momento da matrícula	Numérica
Internacional	Se o aluno é um estudante internacional	Categórica
Unidades curriculares 1º semestre (creditadas)	Número de unidades curriculares creditadas pelo aluno no primeiro semestre.	Numérica
Unidades curriculares 1º semestre (inscrições)	Número de unidades curriculares matriculadas pelo aluno no primeiro semestre.	Numérica
Unidades curriculares 1º semestre (avaliações)	Número de unidades curriculares avaliadas pelo aluno no primeiro semestre.	Numérica
Unidades curriculares 1º semestre (aprovações)	Número de unidades curriculares aprovadas pelo aluno no primeiro semestre.	Numérica
Nota nas Unidades curriculares 1º semestre	Nota nas Unidades curriculares 1º semestre	Numérica
Unidades Curriculares 1º Semestre (sem Avaliações)	Unidades curriculares 1º semestre (sem avaliações)	Numérica
Unidades curriculares 2º semestre (creditadas)	Número de unidades curriculares creditadas pelo aluno no segundo semestre.	Numérica
Unidades curriculares 2º semestre (inscrições)	Número de unidades curriculares matriculadas pelo aluno no segundo semestre.	Numérica
Unidades curriculares 2º semestre (avaliações)	Número de unidades curriculares avaliadas pelo aluno no segundo semestre.	Numérica
Unidades curriculares 2º semestre (aprovações)	Número de unidades curriculares aprovadas pelo aluno no segundo semestre.	Numérica
Nota nas Unidades curriculares 2º semestre	Nota nas Unidades curriculares segundo semestre	Numérica
Unidades Curriculares 2º Semestre (sem Avaliações)	Unidades curriculares segundo semestre (sem avaliações)	Numérica
Taxa de Desemprego	Taxa de Desemprego	Numérica
Taxa de inflação	Taxa de inflação	Numérica
PIB	Produto interno Bruto	Numérica
Target	variável de resposta	Numérica

Tabela 3 – Atributos da base de dados

Fonte: Elaborado pelo Autor

O gráfico figura 16 representa o atributo alvo (target) nesta distribuição dos alunos em diferentes situações acadêmicas, sendo parte de uma análise para identificar possíveis causas de evasão. Dos 4424 alunos analisados, 2209 foram graduados, 794 estão atualmente matriculados e 1421 evadiram. O que pode ser útil para identificar padrões ou tendências relacionadas à evasão.

Levando em conta que o foco da pesquisa são os alunos (evadidos e graduados), optou-se pela retirada das informações referentes aos matriculados, uma vez que não permite o uso de métricas por não ter informação se estes alunos se formaram ou desistiram do curso.

O gráfico 17 apresenta a distribuição da média de idade dos alunos no momento da matrícula. Observa-se que a maioria dos alunos tem idades entre 18 e 22 anos, com um pico em torno da idade 60 anos. Essa representação visual fornece *insights* sobre a faixa etária dos alunos no momento da matrícula, o que pode ser útil para entender a composição etária da população estudantil e identificar possíveis correlações entre a idade e a evasão.

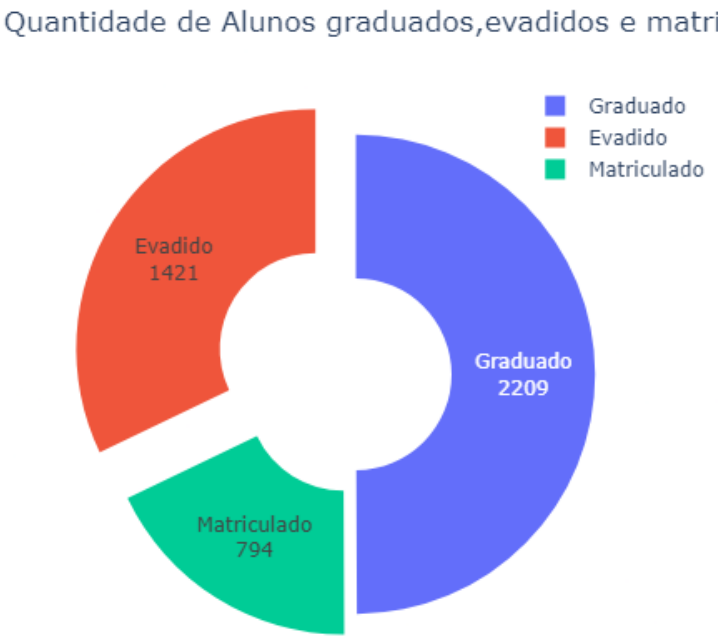

Figura 16 – Gráfico da Variável de Resultado da base dados adotada para este trabalho

Fonte: Elaborado pelo Autor

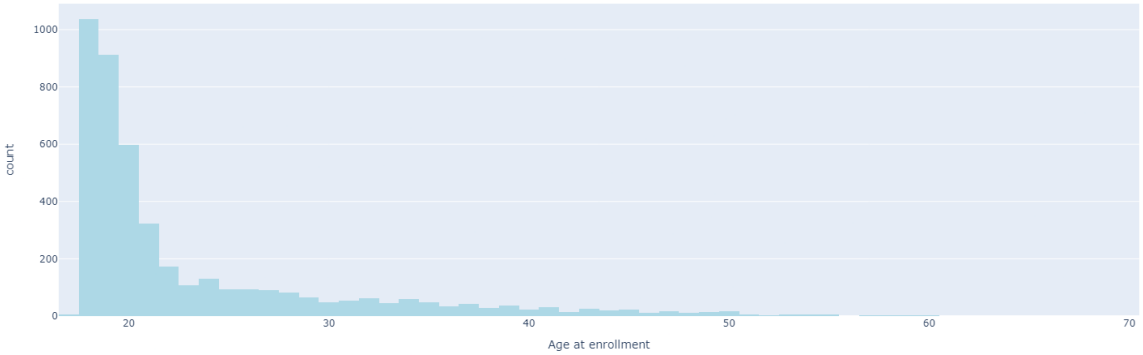

Figura 17 – Gráfico de Distribuição da idade dos alunos no momento da matrícula

Fonte: Elaborado pelo Autor

4.4 Metodologias de Aprendizado de Máquina

Conforme a informação sobre evasão de alunos, que se apresenta essencialmente como uma classificação binária, com "*0*"para formado e "*1*"para evasão/abandono, a escolha dos algoritmos de aprendizado de máquina foi baseada na necessidade de explorar diferentes técnicas de aprendizado supervisionado. O objetivo é identificar qual deles oferece o melhor desempenho na tarefa de prever a evasão de alunos, visando uma tomada de decisão segura e eficaz.

A variedade de algoritmos de aprendizado de máquina escolhidos possibilitou uma comparação abrangente, levando à identificação do modelo mais apropriado para o conjunto de dados específico em questão.

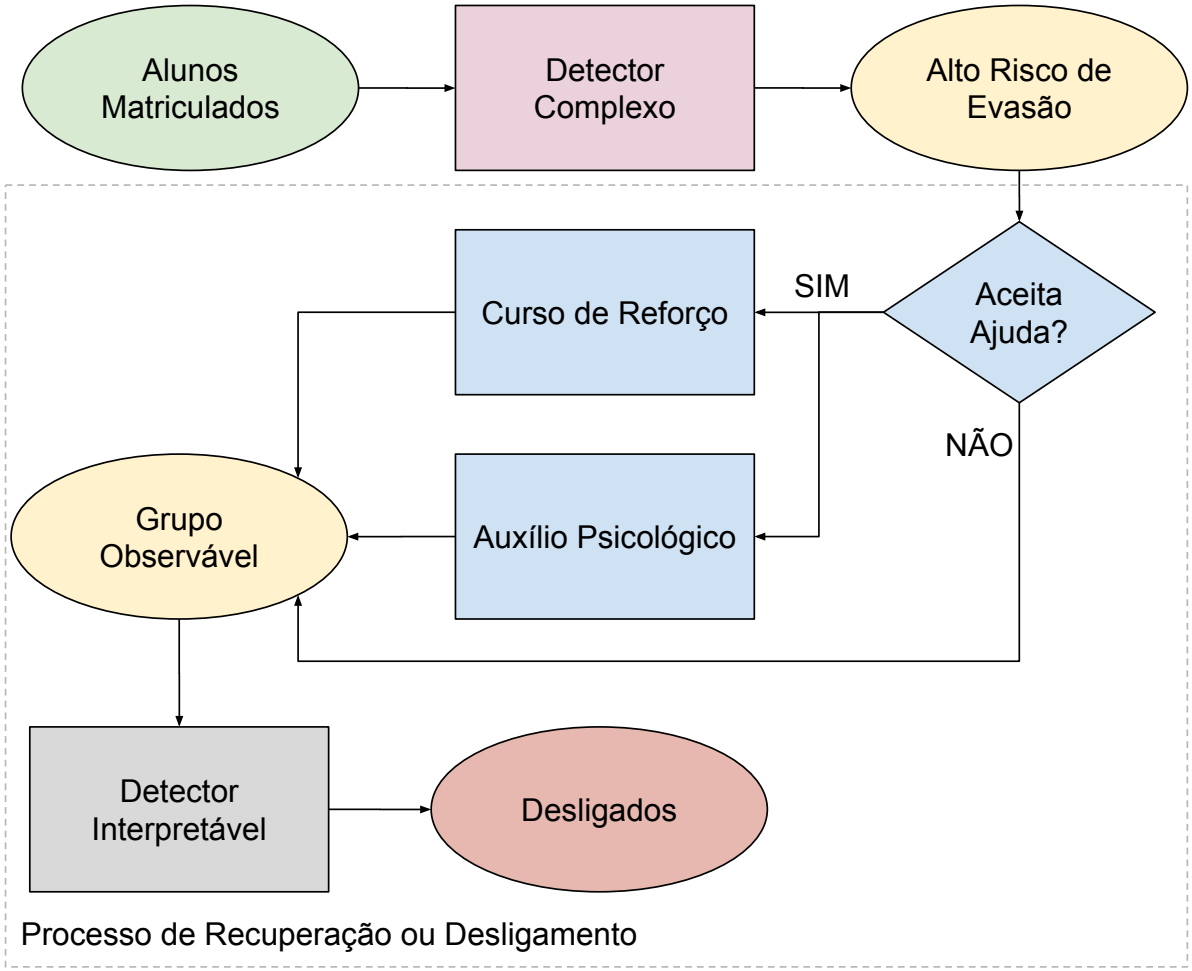

Figura 18 – Procedimento proposto para recuperação ou desligamento do aluno

Fonte: Elaborado pelo Autor

Conforme apresentado na figura 18 os modelos de estudo foram categorizados em dois grupos: interpretáveis e complexos.

4.4.1 Algoritmos Interpretáveis

Interpretáveis são algoritmos que possuem a característica de serem passíveis de validação jurídica, uma vez que possibilitam a demonstração, análise e interpretação dos resultados de maneira transparente. Esses algoritmos permitem a criação de modelos que viabilizam a identificação de alunos que possivelmente evadirão da instituição.

A transparência e a capacidade de interpretação desses algoritmos são fundamentais para garantir a justiça e eficácia das intervenções no ambiente educacional.

Abaixo a lista de modelos interpretáveis:

1. *DecisionTreeClassifier*
2. *ExtraTreesClassifier*
3. *GradientBoostingClassifier*
4. *RandomForestClassifier*

4.4.2 Algoritmos Complexos

Algoritmos complexos são aqueles em que os resultados detalhados não podem ser facilmente interpretados. Suas variáveis e processos internos são de difícil compreensão, o que torna desafiador identificar o motivo específico por trás das decisões, como a evasão ou permanência do aluno. Essa falta de transparência pode dificultar a análise e a interpretação dos resultados, limitando a capacidade de compreender plenamente o funcionamento do algoritmo em questão. Deste modo, este modelo visa apresentar os alunos com potencial para se beneficiar de reforço escolar ou suporte psicológico/emocional, buscando assim também a prevenção da evasão escolar.

Os algoritmos complexos são:

1. *RidgeClassifier*
2. *LogisticRegression*
3. *DummyClassifier*
4. *LinearDiscriminantAnalysis*
5. *QuadraticDiscriminantAnalysis*
6. *AdaBoostClassifier*
7. *BaggingClassifier*
8. *HistGradientBoostingClassifier*
9. *GaussianProcessClassifier*
10. *BernoulliNB*
11. *GaussianNB*
12. *KNeighborsClassifier*
13. *NearestCentroid*

14. *MLPClassifier*
15. *LinearSVC*
16. *NuSVC*
17. *SVC*
18. *XGBRFClassifier*
19. *XGBClassifier*
20. *LGBMClassifier*

4.4.3 Necessidades Jurídicas

A interpretabilidade do modelo de aprendizado de máquina para prever a evasão escolar é importante do ponto de vista jurídico. A transparência e a capacidade de explicar como o modelo toma suas decisões são fundamentais para garantir a conformidade com leis e regulamentos relacionados a direito dos alunos.

Em casos de desligamento de alunos com base em previsões feitas por um modelo de aprendizado de máquina, é essencial que as razões por trás dessas previsões sejam compreensíveis e justificáveis. A interpretabilidade do modelo ajuda a garantir que não haja discriminação injusta ou viés nas decisões tomadas, o que é essencial para cumprir com leis antidiscriminatórias.

Além disso, a capacidade de interpretar o modelo permite que os responsáveis pela tomada de decisão compreendam melhor quais variáveis ou características dos alunos estão influenciando as previsões de evasão. Isso pode levar a intervenções mais eficazes e personalizadas para ajudar os alunos em risco, contribuindo para a melhoria do processo de desligamento e para a transparência nas práticas educacionais.

4.5 Treinamento e Validação

No processo de treinamento e validação, a etapa é realizada da seguinte maneira:

1. **Treinamento do Modelo:** Em cada iteração do *k-fold*, o modelo é treinado utilizando k-1 *folds* como conjunto de treinamento. Isso significa que o modelo é ajustado para aprender os padrões e relações nos dados disponíveis, com o objetivo de fazer previsões precisas.

2. **Validação do Modelo:** Após o treinamento, o modelo é avaliado no *fold* restante que foi retido como conjunto de validação. Nesta etapa, o modelo faz previsões com

base nos dados do *fold* de validação e as métricas de desempenho são calculadas para avaliar quão bem o modelo generaliza para novos dados.

3. **Avaliação do Desempenho:** As métricas de desempenho, como acurácia, precisão, recall, entre outras, são registradas para cada *fold* de validação. Essas métricas ajudam a avaliar a eficácia do modelo em fazer previsões precisas e identificar possíveis áreas de melhoria.

4. **Iteração:** O processo de treinamento e validação é repetido k vezes, cada vez utilizando um *fold* diferente como conjunto de validação. Isso garante que o modelo seja avaliado em diferentes conjuntos de dados e ajuda a reduzir o viés na avaliação do desempenho.

Ao final das k iterações, as métricas de desempenho obtidas em cada *fold* de validação são combinadas para fornecer uma estimativa mais confiável do desempenho geral do modelo. Esse processo ajuda a garantir que o modelo seja robusto e capaz de generalizar bem para novos dados.

4.6 Implementação e Aprimoramento

Após completar as etapas de treinamento, validação e avaliação do modelo de aprendizado de máquina utilizando a validação cruzada k-fold, o processo de implementação e aprimoramento envolve a integração do modelo em um ambiente de produção, o monitoramento contínuo do desempenho, a identificação de áreas de melhoria e ajustes com base nas considerações recebidas dos usuários. É importante reavaliar periodicamente o modelo utilizando novos dados e repetir o processo de validação cruzada para garantir que o modelo continue a ser eficaz e preciso ao longo do tempo.

O código gerado de todas as implementações da pesquisa está disponível no repositório (MALERBA; MORAES, 2024).

4.7 Modelo de ML SIGAA - UFRN

O SIGAA (Sistema Integrado de Gestão de Atividades Acadêmicas) desenvolvido e mantido pela UFRN (Universidade Federal do Rio Grande do Norte) desempenha um papel de extrema importância na gestão e organização das atividades acadêmicas em diversas instituições de ensino superior.

No âmbito deste sistema, foi desenvolvido um modelo de aprendizado de máquina (figura 19) com o objetivo de analisar e prever possíveis casos de evasão de alunos. Os resultados iniciais foram promissores, revelando significativas informações que poderiam contribuir para a melhoria da retenção e do acompanhamento dos estudantes.

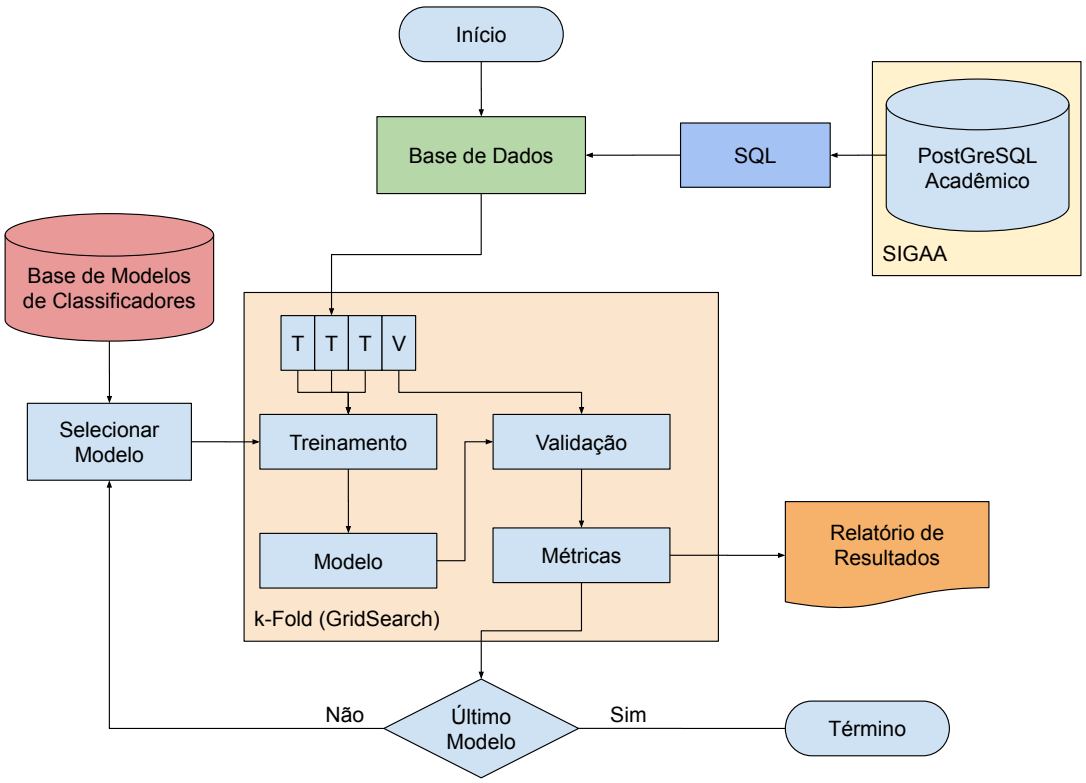

Figura 19 – Modelo ML Sigaa UFRN

Fonte: Elaborado pelo Autor

Na figura 20 é demonstrado o modelo de Entidade e Relacionamento, no qual podemos começar com a entidade "Pessoa", que tem atributos como ID_Pessoa, nome, data_nascimento,estado_civil e outros atributos pessoais. A relação entre Pessoa e Discente pode ser representada por uma relação "é um"ou "é aluno de". A entidade Discente pode ter atributos como Id_discente,id_pessoa, matrícula, e outros dados específicos do aluno.

Além disso, a entidade "Curso"pode ter atributos como ID_Curso, nome_curso, carga_horaria, entre outros. A relação entre Discente e Curso pode ser representada por uma relação "está matriculado em"ou "faz parte de".

A forma de ingresso e o status do discente podem ser atributos da entidade Discente. O município também pode ser um atributo da entidade Pessoa.

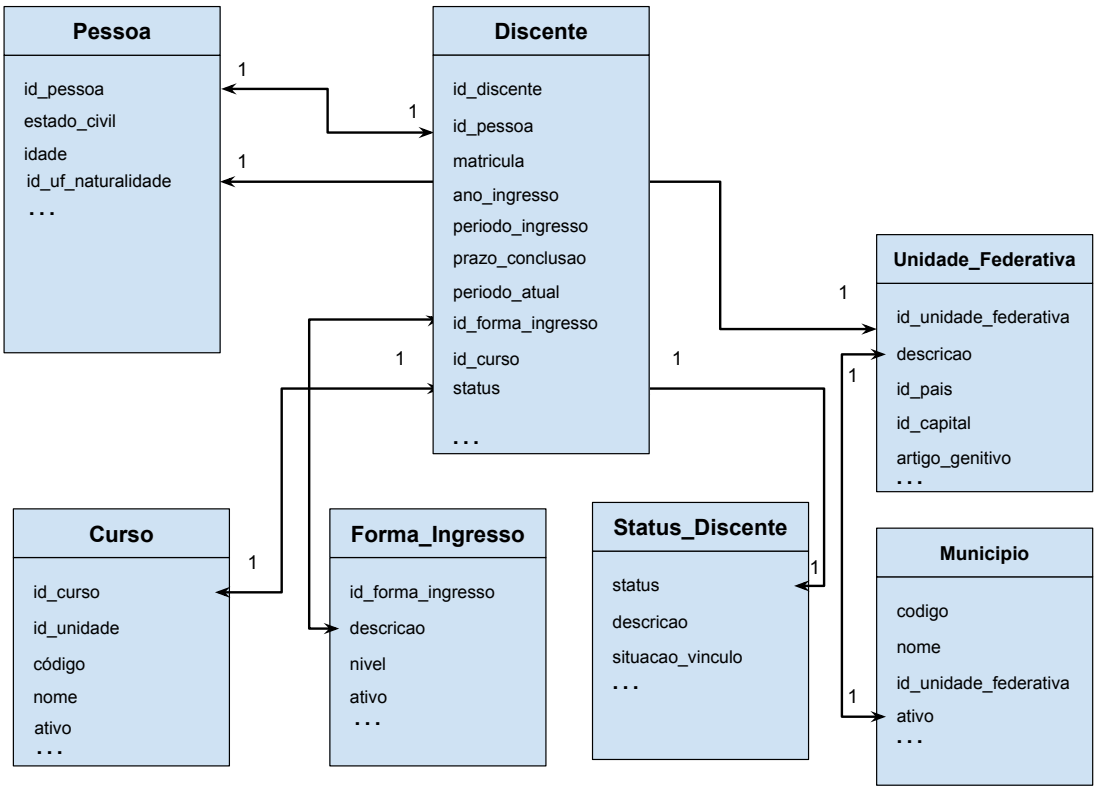

Figura 20 – Diagrama de Entidade e Relacionamento - Sigaa

Fonte: Elaborado pelo Autor

5 Resultados

Neste capítulo, serão apresentados os resultados obtidos a partir da aplicação dos algoritmos complexos e interpretáveis no modelo de previsão de evasão de alunos.

5.1 Modelo para Recuperação de possíveis alunos evadidos

O Modelo para Recuperação de possíveis alunos evadidos é uma abordagem estratégica e sistemática para identificar e apoiar estudantes que estejam em risco de evasão acadêmica. Esse modelo visa prevenir a evasão, promovendo a retenção dos alunos e garantindo que recebam o suporte necessário para superar desafios acadêmicos, pessoais ou emocionais que possam estar impactando seu desempenho. Através da análise de indicadores de alerta precoce e da implementação de intervenções personalizadas, o Modelo para Recuperação busca reverter a tendência de evasão, contribuindo para o sucesso e a conclusão dos estudos dos alunos.

Conforme tabela 4, o objetivo foi encontrar os melhores resultados para F1-score de Evasão dos alunos. Para este modelo os resultados dos algoritmos MLPClassifier,LogisticRegression e LinerarSVC foram os que alcançaram os melhores resultados.

Os métodos (MaxAbsScaler,StandardScaler,RobustScaler) foi utilizado para minimizar as variações excessivas dos dados.

O modelo LogisticRegression com StandardScaler obteve uma pontuação de F1 de 93,18% para a previsão de formandos e 88,55% para evasão. Além disso, alcançou uma acurácia de 95,97% para formandos e 84,44% para evasão, com Tempo de treinamento médio: 0.013533 segundos e Tamanho do modelo: 1033 bytes. O modelo de regressão logística com padronização dos dados obteve um bom equilíbrio entre precisão e recall para ambas as classes.

O HistGradientBoostingClassifier com StandardScaler apresentou uma pontuação de F1 de 92,81% para formandos e 87,82% para evasão, com acurácia de 95,83% e 83,38%, com tempo de treinamento médio: 0.077165 segundos e tamanho do modelo: 158611 bytes.O modelo de boosting com padronização dos dados também apresentou um bom desempenho, com alta acurácia e F1-score.

O XGBClassifier com StandardScaler obteve uma pontuação de F1 de 92,48% para formandos e 87,36% para evasão, com acurácia de 95,29% e 83,25%, com tempo de treinamento médio: 0.040014 segundos e tamanho do modelo: 228645 bytes. O modelo XGBoost com padronização também teve um bom desempenho, embora ligeiramente inferior aos dois modelos anteriores.

Em resumo, os modelos de regressão logística, gradient boosting e XGBoost com padronização dos dados são os mais promissores com base nas métricas apresentadas. A importância da escolha adequada do modelo e da técnica de pré-processamento é destacada para melhorar a precisão das previsões e a eficácia do sistema de classificação.

modelo	pre	f1_formando	f1_evasao	ac_formando	ac_evasao	fit_time	size
MLPClassifier	MaxAbsScaler	0,930446714	0,885596334	0,950211601	0,85640073	1,036198298	99555
LogisticRegression	StandardScaler	0,931880074	0,885567603	0,959735652	0,844446945	0,026041746	1033
LogisticRegression	RobustScaler	0,931742909	0,884958759	0,96063905	0,842323576	0,015630404	1033
LinearSVC	StandardScaler	0,9307451	0,881829344	0,964247556	0,832545747	0,080471595	919
MLPClassifier	MinMaxScaler	0,927614971	0,880724883	0,947950181	0,850857242	1,292704821	99555
LinearSVC	MaxAbsScaler	0,930496482	0,880522647	0,966510247	0,827558978	0,026772102	919
LinearSVC	MinMaxScaler	0,930496482	0,880522647	0,966510247	0,827558978	0,018970887	918
HistGradientBoostingClassifier	StandardScaler	0,928150115	0,878266023	0,958387396	0,83385696	0,080099662	158611
HistGradientBoostingClassifier	MinMaxScaler	0,928150115	0,878266023	0,958387396	0,83385696	0,083001852	158611
HistGradientBoostingClassifier	RobustScaler	0,928150115	0,878266023	0,958387396	0,83385696	0,072779973	158611
HistGradientBoostingClassifier	MaxAbsScaler	0,928150115	0,878266023	0,958387396	0,83385696	0,077658494	158611
HistGradientBoostingClassifier	NoneType	0,928150115	0,878266023	0,958387396	0,83385696	0,078837474	158611
MLPClassifier	RobustScaler	0,925250332	0,876545606	0,947049325	0,844509113	1,106666962	99555
XGBClassifier	StandardScaler	0,924889327	0,87364124	0,952975807	0,832506868	0,040018797	228645
XGBClassifier	NoneType	0,924889327	0,87364124	0,952975807	0,832506868	0,051207383	228645
XGBClassifier	MaxAbsScaler	0,924889327	0,87364124	0,952975807	0,832506868	0,040653149	228645
XGBClassifier	MinMaxScaler	0,924889327	0,87364124	0,952975807	0,832506868	0,254612128	228645
XGBClassifier	RobustScaler	0,924889327	0,87364124	0,952975807	0,832506868	0,04106458	228645
SGDClassifier	MinMaxScaler	0,925034818	0,873593971	0,952027593	0,83460713	0,005811135	1276
LinearSVC	RobustScaler	0,92562567	0,873532249	0,957941165	0,82621809	0,065876166	919
LogisticRegression	MinMaxScaler	0,92478661	0,869884365	0,962434608	0,814878494	0,01136597	1033
LogisticRegression	MaxAbsScaler	0,92455149	0,869555971	0,96197617	0,814878494	0,010329644	1033
RidgeClassifier	MinMaxScaler	0,926832855	0,868804902	0,977388809	0,795290344	0,001853784	1035
RidgeClassifier	MaxAbsScaler	0,926832855	0,868804902	0,977388809	0,795290344	0,001970291	1035
RidgeClassifier	RobustScaler	0,925608985	0,867370731	0,974219008	0,796717818	0,001855532	1035
LinearDiscriminantAnalysis	StandardScaler	0,925314852	0,867237457	0,972867039	0,798118434	0,008986473	2299
LinearDiscriminantAnalysis	MinMaxScaler	0,925314852	0,867237457	0,972867039	0,798118434	0,008672078	2299
LinearDiscriminantAnalysis	MaxAbsScaler	0,925314852	0,867237457	0,972867039	0,798118434	0,018713713	2299
LinearDiscriminantAnalysis	RobustScaler	0,925314852	0,867237457	0,972867039	0,798118434	0,009029547	2299

Tabela 4 – Modelo para Desligamento do Aluno

5.2 Modelo para Desligamento de Aluno

Um Modelo para Desligamento de Aluno é uma abordagem estratégica e organizada para gerenciar e formalizar o desligamento de estudantes de uma instituição educacional. Esse modelo visa estabelecer procedimentos claros e transparentes para lidar com situações em que um aluno precisa ser desligado, seja por motivos acadêmicos, disciplinares ou pessoais. Ao definir critérios e processos para o desligamento, o Modelo para Desligamento de Aluno busca garantir a equidade, a consistência e o respeito no tratamento dos casos de desligamento, assegurando que a decisão seja tomada de forma justa e em conformidade com as políticas institucionais.

Para este o foco principal foi a métrica F1-score para Formandos, visando minimizar ao máximo os erros ao identificar alunos com potencial para formar que possam estar na lista de desligamento.

Os algoritmos, como RandomForestClassifier, GradientBoostingClassifier, DecisionTreeClassifier, ExtraTreesClassifier e ExtraTreeClassifier, foram avaliados quanto à sua

eficácia na identificação de padrões de evasão de alunos.

Os resultados revelaram diferentes desempenhos em termos de F1-score, acurácia e eficiência computacional. Por exemplo, o RandomForestClassifier demonstrou um bom equilíbrio entre precisão e recall, enquanto o GradientBoostingClassifier obteve alta acurácia, mas com um tempo de ajuste mais elevado, como demonstrado na tabela 5.

1. **RandomForestClassifier:** Alcançou um F1-score de 91,44% para formandos e 84,20% para evasão, com acurácia de 97,42% para formandos e 75,66% para evasão, com tempo médio de treinamento de 0,0043 segundos e tamanho do modelo de 11130 bytes. O modelo Random Forest obteve alta acurácia, mas o F1-score para evasão é um pouco menor. Ele é robusto e pode ser uma boa escolha para classificação.

2. **GradientBoostingClassifier:** Apresentou resultados sólidos com um F1-score de 91,35% para formandos e 83,84% para evasão, e acurácia de 97,74% para formandos e 74,75% para evasão, com tempo médio de treinamento de 0,0128 segundos e tamanho do modelo de 13010 bytes. O modelo Gradient Boosting tem alta acurácia, mas o F1-score para evasão é um pouco menor. Ele é eficaz para previsões, mas pode ser mais lento no treinamento.

3. **DecisionTreeClassifier:** Obteve um F1-score de 91,31% para formandos e 84,53% para evasão, com acurácia de 96,06% para formandos e 77,69% para evasão, com tempo médio de treinamento de 0,0040 segundos e tamanho do modelo de 3185 bytes. A árvore de decisão tem alta acurácia e F1-score. É um modelo simples e rápido de treinar.

4. **ExtraTreesClassifier:** Alcançou um F1-score de 88,36% para formandos e 77,42% para evasão, com acurácia de 95,39% para formandos e 67,98% para evasão, com tempo médio de treinamento de 0,0034 segundos e tamanho do modelo: 10120 bytes. O Extra Trees é semelhante ao Random Forest, mas com maior variância. Pode ser útil quando a diversidade é necessária.

5. **ExtraTreeClassifier:** Apresentou um F1-score de 81,61% para formandos e 57,04% para evasão, com acurácia de 93,28% para formandos e 44,94% para evasão, com tempo médio de treinamento de 0,0006 segundos e tamanho do modelo de 2902 bytes. O Extra Tree é simples e rápido, mas o F1-score para evasão é baixo. Pode ser menos robusto.

O Random Forest é uma escolha sólida com alta acurácia e bom equilíbrio entre F1-score para formandos e evasão. O Decision Tree também é uma opção rápida e eficaz.

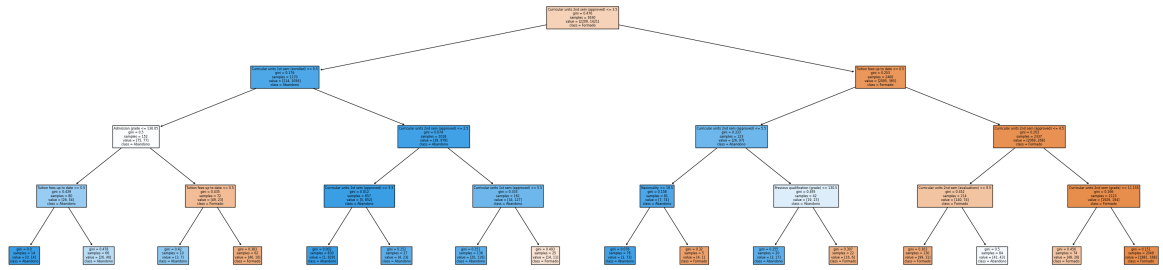

Figura 21 – Acertos Formados: 92.13% Abandonos: 86.59%

Fonte: Elaborado pelo Autor

modelo	pre	f1_formando	f1_evasao	ac_formando	ac_evasao	fit_time	size
RandomForestClassifier	NoneType	0,914409088	0,842044073	0,974202989	0,756613852	0,006240129	11130
GradientBoostingClassifier	NoneType	0,913502404	0,838373666	0,977396334	0,747451718	0,018120845	12975
DecisionTreeClassifier	NoneType	0,91308356	0,845278975	0,960624401	0,776927949	0,004046122	3185
ExtraTreesClassifier	NoneType	0,883647404	0,774249722	0,953868167	0,679770441	0,005463044	10120
ExtraTreeClassifier	NoneType	0,81611676	0,570411523	0,932763005	0,449432015	0,000867367	2902

Tabela 5 – Modelos para Desligamento de Aluno

Para abordar a preocupação com a questão jurídica relacionada ao desligamento de alunos, é fundamental considerar as políticas e regulamentos educacionais vigentes. O desligamento de alunos pode envolver questões legais sensíveis, como direitos dos estudantes, legislação educacional, contratos ou acordos acadêmicos, entre outros.

Ao criar um gráfico demonstrativo de uma árvore de decisão para representar as etapas do processo de desligamento de alunos, é possível visualizar de forma clara e hierárquica as diferentes decisões e critérios que levam ao desligamento de um aluno, respeitando os aspectos legais e regulatórios.

A árvore de decisão mostra como o modelo toma decisões com base nas características dos dados. Cada nó representa uma decisão com base em uma característica específica. Os rótulos nas folhas indicam a classe prevista (Formado ou Abandono).

Além disso, para representar as regras de forma descritiva, pode-se elaborar um documento detalhado que descreva os procedimentos, critérios e justificativas para o desligamento de alunos, seguindo as diretrizes legais e normativas pertinentes. Esse documento pode incluir informações sobre os motivos para o desligamento, os passos a serem seguidos, os direitos dos alunos e as obrigações da instituição de ensino.

Ao adotar essas abordagens visuais e descritivas, é possível garantir que o processo de desligamento de alunos seja transparente, justo e esteja em conformidade com as leis e regulamentos educacionais aplicáveis.

Na Figura 21 representação visual de uma árvore de decisão, gerada do modelo de aprendizado de máquina para desligamento de Aluno.

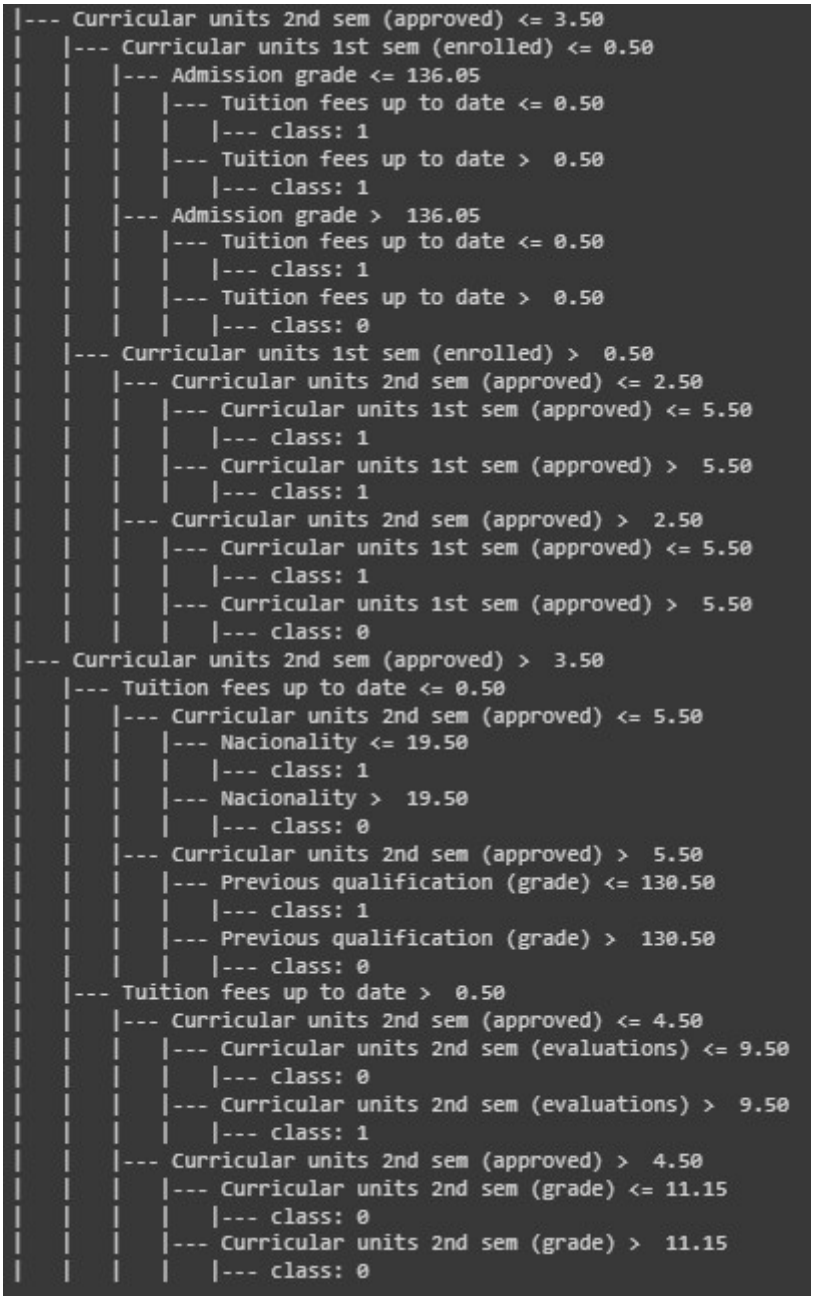

Figura 22 – Forma Descritiva da Arvore de Decisão

Fonte: Elaborado pelo Autor

Na Figura 22 foi representado as regras de forma descritiva.

6 Conclusão

Ao longo desta dissertação, exploramos a aplicação de diversos algoritmos de aprendizado de máquina, tanto complexos quanto interpretáveis, na previsão de evasão de alunos. Os resultados obtidos revelaram números valiosos sobre a eficácia desses modelos na identificação de alunos em risco de evasão e na tomada de decisões para intervenções educacionais.

Os algoritmos complexos demonstraram um bom desempenho em termos de métricas de avaliação, enquanto os algoritmos interpretáveis proporcionaram uma compreensão mais clara das razões por trás das previsões, permitindo uma abordagem mais transparente e interpretável.

Com base nos resultados obtidos, é possível concluir que a combinação de diferentes tipos de algoritmos pode ser benéfica para uma abordagem abrangente na previsão da evasão escolar. Esses resultados podem contribuir significativamente para a gestão educacional, auxiliando na identificação precoce de alunos em risco de evasão e na implementação de estratégias eficazes para a retenção e o sucesso dos estudantes.

Em relação aos objetivos alcançados, podemos destacar os seguintes pontos :

1. **Busca Bibliográfica e Teorias Relacionadas:**

   Foi realizada uma extensa busca na literatura existente para identificar as possíveis causas da evasão escolar. Isso nos permitiu embasar nosso estudo em teorias e modelos já estabelecidos. A compreensão das causas subjacentes é fundamental para desenvolver modelos eficazes de previsão.

2. **Coleta e Processamento de Dados:**

   Foram coletados os dados dos alunos e do ambiente acadêmico, incluindo informações demográficas, socioeconômicas e acadêmicas. O processamento desses dados foi essencial para prepará-los para a construção dos modelos de aprendizado de máquina.

3. **Treinamento de Modelos:**

   Foram desenvolvidos e treinados diferentes modelos de aprendizado de máquina. Alguns deles incluíram redes neurais, árvores de decisão e regressão logística. O foco era identificar os alunos em risco de evasão com base nos dados disponíveis e assim dar a opção de se recuperarem antes de um possível desligamento.

4. **Teste e Validação dos Modelos:**

Utilizou-se a validação cruzada k-fold para avaliar o desempenho dos modelos. A validação rigorosa permitiu verificar a capacidade de generalização dos modelos.

5. **Análise dos Resultados:** Analisou-se as métricas de desempenho, como acurácia, precisão e sensibilidade.

Os modelos demonstraram ser promissores na identificação precoce de alunos em risco.

Sendo assim, conclui-se que este estudo destaca a relevância do aprendizado de máquina no contexto educacional e reforça o potencial dessas ferramentas para melhorar a qualidade da educação, promovendo um ambiente escolar mais eficaz e inclusivo.

6.1 Trabalhos Futuros

6.1.1 Aprimoramento do Sistema SIGAA e Exploração de Dados de Evasão de Alunos

Uma área promissora para pesquisas futuras envolve o estudo aprofundado no uso do sistema SIGAA para prevenção de evasão de alunos, com destaque para a aplicabilidade em análise de dados de evasão de alunos. Algumas direções potenciais para investigações futuras incluem:

1. **Otimização do Sistema SIGAA**: Explorar oportunidades para otimizar o desempenho e a usabilidade do sistema SIGAA, visando uma melhor experiência para os usuários e uma gestão mais eficiente dos dados acadêmicos.

2. **Integração de Novas Funcionalidades**: Pesquisar a possibilidade de integrar novas funcionalidades ao SIGAA, como análises avançadas de dados, modelos preditivos de evasão e ferramentas de suporte à decisão para auxiliar na identificação precoce de alunos em risco.

3. **Validação e Generalização dos Resultados**: Realizar estudos adicionais para validar e generalizar os resultados obtidos com o sistema SIGAA em relação a evasão de alunos, ampliando assim a aplicabilidade e relevância do sistema em diferentes contextos educacionais.

Referências

ALGHAMDI, M. et al. Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 7, p. e0179805, 2017. Citado na página 31.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 27.

AMARAL, F. *Introdução à ciência de dados: mineração de dados e big data*. [S.l.]: Alta Books Editora, 2016. Citado na página 35.

ASSIS, L. R. S. d. Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados. 2018. Citado na página 41.

ASTIN, A. W. Student involvement: A developmental theory for higher education. ACPA Executive Office, 1999. Citado na página 19.

AULCK, L. et al. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016. Citado na página 40.

BALANIUK, R. et al. Predicting evasion candidates in higher education institutions. In: SPRINGER. *Model and Data Engineering: First International Conference, MEDI 2011, Óbidos, Portugal, September 28-30, 2011. Proceedings 1*. [S.l.], 2011. p. 143–151. Citado na página 40.

BARBOSA, D. 14 causas do abandono escolar no brasil. 2017. Acessado: 10 Dez. 2023. Disponível em: <https://www.politize.com.br/abandono-escolar-causas/>. Acesso em: 10 Dez. 2023. Citado na página 25.

BARBOSA, D. causas do abandono escolar no brasil. 2017. *acesso em*, v. 20, 2021. Citado 2 vezes nas páginas 24 e 42.

BELLHOUSE, D. R. The reverend thomas bayes, frs: a biography to celebrate the tercentenary of his birth. 2004. Citado na página 34.

BIAMONTE, J. et al. Quantum machine learning. *Nature*, Nature Publishing Group UK London, v. 549, n. 7671, p. 195–202, 2017. Citado na página 16.

BIASI, F. D. *Utilização dos Algoritmos Naïve Bayes e Decision Tree no Auxílio ao Diagnóstico da Doença de Alzheimer por meio do Processamento de Textos Transcritos de Discurso de Pacientes*. 2019. <https://sapiens.ipt.br/Teses/2019_EC_Paulo_Biasi.pdf>. [Accessed 23-01-2024]. Citado na página 34.

BISSOLI, A. C. d. S.; RODRIGUES, R. M. I. *Evasão escolar: o caso do Colégio Estadual Antonio Francisco Lisboa, 2010*. 2017. Citado na página 19.

BRASIL. Fies: Prestação de contas ordinárias anual - relatório de gestão do exercício de 2014. p. http://portal.mec.gov.br/index.php?option=com$_docman$&$view$ = $download$&$alias$ = 66631 − −$relatorio$ − −$gesta$ − −$Fies$ − −$exercicio$ − −2016 − −pdf&$category_slug$ = $junho$ − −2017 − −pdf&$Itemid$ = 3019, 2015. $Citadonapágina16$.

CAMPOS, J. D. d. S. *Fatores explicativos para a evasão no Ensino Superior através da análise de sobrevivência: o caso da UFPE*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2016. Citado na página 26.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm New York, NY, USA, v. 2, n. 3, p. 1–27, 2011. Citado na página 33.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, p. 273–297, 1995. Citado na página 33.

COSTA, A. G. d. *Aplicação de técnicas de mineração de dados e Learning Analytics para predição de evasão de alunos nos cursos de Ciência da Computação e Engenharias da UFPel*. Dissertação (Mestrado) — Universidade Federal de Pelotas, 2021. Citado 2 vezes nas páginas 37 e 42.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. In: *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain*. [S.l.: s.n.], 2009. p. 41–50. Citado na página 39.

DELEN, D. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, SAGE Publications Sage CA: Los Angeles, CA, v. 13, n. 1, p. 17–35, 2011. Citado 2 vezes nas páginas 30 e 40.

ENAP. Análise de dados em linguagem r. 2020. Acessado: 15 Nov. 2023. Disponível em: <https://www.escolavirtual.gov.br/curso/325>. Acesso em: 15 Nov. 2023. Citado na página 28.

ESCOVEDO, T.; KOSHIYAMA, A. *Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise*. [S.l.]: Casa do Código, 2020. Citado 5 vezes nas páginas 27, 30, 31, 34 e 35.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2021. Citado na página 35.

FARIA, M. M. Detecção de intrusões em redes de computadores com base nos algoritmos knn, k-means++ e j48. *São Paulo: Dissertação (Programa de Mestrado em Ciência da Computação)—Faculdade Campo Limpo Paulista–FACCAMP*, 2016. Citado na página 35.

FERREIRA, J. A. B. Redes neurais artificiais aplicadas em aprendizagem de trajetória em robótica móvel. 2020. Citado na página 36.

FIALHO, M. G. D. et al. A evasão escolar e a gestão universitária: o caso da universidade federal da paraíba. Universidade Federal da Paraíba, 2014. Citado na página 19.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, p. 641–659, 2007. Citado 3 vezes nas páginas 15, 19 e 26.

FRANK, E.; HALL, M. A.; WITTEN, I. H. *The WEKA workbench*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 15.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. [S.l.]: Alta Books, 2019. Citado 2 vezes nas páginas 33 e 36.

GIL, A. C. et al. *Como elaborar projetos de pesquisa*. [S.l.]: Atlas São Paulo, 2002. v. 4. Citado na página 43.

GRUS, J. *Data science do zero*. [S.l.: s.n.], 2016. Citado na página 31.

HAN, J.; KAMBER, M.; PEI, J. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012. Citado na página 30.

HASTIE, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2. Citado na página 30.

IDEB. Índice de desenvolvimento da educação básica (ideb): metas intermediárias para a sua trajetória no brasil, estados, municípios e escolas. *Brasil: INEP/MEC*, 2007. Citado na página 20.

INEP, I. N. de Estudos e P. E. A. T. Sinopse estatística da educação básica. *Ministério da Educação*, 1998. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/sinopse_estatistica_da_educacao_basica_censo_escolar_98.pdf">. Citado na página 20.

JADRIĆ, M.; GARAČA, Ž.; ČUKUŠIĆ, M. Student dropout analysis with application of data mining methods. *Management: journal of contemporary management issues*, Sveučilište u Splitu, Ekonomski fakultet, v. 15, n. 1, p. 31–46, 2010. Citado na página 39.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado na página 30.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2018. Citado na página 35.

KAGGLE. *Predict students' dropout and academic success*. 2021. Acessado: 13 Dez. 2023. Disponível em: <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention?resource=download>. Acesso em: 13 Dez. 2023. Citado na página 47.

KARAMOUZIS, S. T.; VRETTOS, A. An artificial neural network for predicting student graduation outcomes. In: *Proceedings of the World congress on engineering and computer science*. [S.l.: s.n.], 2008. p. 991–994. Citado na página 39.

LEMOS, Í. V. d. R. *Prevendo a evasão escolar em uma instituição de ensino técnico utilizando mineração de dados educacionais*. Dissertação (B.S. thesis) — Brasil, 2021. Citado 2 vezes nas páginas 27 e 42.

LOPES, S. R. d. A. Evasão de alunos no ensino superior: ações para reduzir a evasão nas ies. 2023. Citado na página 19.

MALERBA, A.; MORAES, C. H. V. *Evasão de Alunos*. [S.l.]: GitHub, 2024. <https://github.com/admalerba/Mestrado>. Citado na página 54.

MANHÃES, L. M. B. et al. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*. [S.l.: s.n.], 2011. v. 1, n. 1. Citado na página 40.

MARIA, W.; DAMIANI, J. L.; PEREIRA, M. Rede bayesiana para previsao de evasao escolar. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2016. v. 5, n. 1, p. 920. Citado na página 41.

MARQUES, F. T. A volta aos estudos dos alunos evadidos do ensino superior brasileiro. *Cadernos de Pesquisa*, SciELO Brasil, v. 50, p. 1061–1077, 2020. Citado 2 vezes nas páginas 25 e 26.

MEDEIROS, N. S. R. d. et al. Análise do desempenho do algortimo k-nearest neighbors na classificação de patologias de coluna vertebral. *Plataforma Espaço Digital*, 2019. Acessado: 23 Dez. 2023. Disponível em: <https://www.editorarealize.com.br/artigo/visualizar/56476>. Acesso em: 13 Dez. 2023. Citado 2 vezes nas páginas 35 e 36.

MELO, A. L. et al. Uso da técnica de mineração de dados como uma ferramenta de gestão da evasão no ensino superior. Universidade Federal do Triângulo Mineiro, 2019. Citado na página 41.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill, 1997. Citado 4 vezes nas páginas 27, 31, 34 e 36.

MOREIRA, F. J. R. Aprendizagem de máquina na predição da evasão no ensino superior. Universidade Federal do Paraná. Setor de Ciências Exatas. Curso de Especialização em Data Science Big Data, 2020. Acessado: 13 Dez. 2023. Disponível em: <https://hdl.handle.net/1884/71062>. Acesso em: 13 Dez. 2023. Citado na página 41.

MUSTAFA, M. N.; CHOWDHURY, L.; KAMAL, M. S. Students dropout prediction for intelligent system from tertiary level in developing country. In: IEEE. *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. [S.l.], 2012. p. 113–118. Citado na página 40.

NASCIMENTO, F. P. d.; SOUSA, F. L. L. Metodologia da pesquisa científica: teoria e prática–como elaborar tcc. *Brasília: Thesaurus*, 2016. Citado na página 43.

OLIVEIRA, A. C. d. Máquina de aprendizagem mínima com opção de rejeição. 2016. Citado na página 35.

OLIVEIRA, F. L. d.; NÓBREGA, L. Evasão escolar: um problema que se perpetua na educação brasileira. *Revista Educação Pública*, v. 21, n. 19, p. 25, 2021. Citado na página 26.

PASCARELLA, E. T.; TERENZINI, P. T. *How college affects students: Findings and insights from twenty years of research*. [S.l.]: ERIC, 1991. Citado na página 16.

PINHEIRO, M. A. L.; SILVA, J. C. da; SOUZA, B. F. de. Aprendizado de máquina aplicado à análise de evasão no ensino superior. *Anais do Computer on the Beach*, p. 512–521, 2018. Citado 2 vezes nas páginas 33 e 41.

PRESTES, E. M. d. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. *Ensaio: Avaliação e Políticas Públicas em Educação*, SciELO Brasil, v. 26, p. 869–889, 2018. Citado 2 vezes nas páginas 21 e 26.

PRIMÃO, A. P. et al. Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no instituto federal de santa catarina. 2022. Citado na página 42.

RESENDE, P. A. A.; DRUMMOND, A. C. A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 3, p. 1–36, 2018. Citado na página 32.

RIFFEL, S. M.; MALACARNE, V. Evasão escolar no ensino médio: o caso do colégio estadual santo agostinho no município de palotina. *O professor PDE e os desafios da escola pública paranaense*, v. 1, p. 01–24, 2010. Citado na página 19.

RODRIGUEZ, J. D.; PEREZ, A.; LOZANO, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 3, p. 569–575, 2009. Citado na página 29.

ROLIM, V. B.; CORDEIRO, F. R.; FERREIRA, R. Reconhecimento de padrões aplicados a comentários de fóruns educacionais. *Centro de Informática, UFPE, João Pessoa*, 2014. Citado na página 32.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence a modern approach*. [S.l.]: London, 2010. Citado na página 31.

SÁ, J. M. d. C. de et al. Análise de crédito utilizando uma abordagem de mineração de dados. *Revista de Engenharia e Pesquisa Aplicada*, v. 3, n. 3, 2018. Citado na página 35.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 49, p. 337–373, 2019. Citado na página 26.

SEMESP, I. Mapa do ensino superior. evasão. 11ª edição. 2021. Acessado: 20 Dez. 2023. Disponível em: <https://www.semesp.org.br/mapa/edicao-11/brasil/evasao/>. Acesso em: 20 Dez. 2023. Citado 2 vezes nas páginas 21 e 22.

SEMESP, I. Mapa do ensino superior. 13ª edição. 2023. Acessado: 20 Dez. 2023. Disponível em: <https://www.semesp.org.br/mapa/edicao-13/brasil/>. Acesso em: 20 Dez. 2023. Citado 2 vezes nas páginas 21 e 23.

SILVA, E. C. R. Relação da evasão escolar com as práticas docentes: um estudo de caso exploratório em uma instituição do ensino superior. Universidade Federal de Itajubá, 2022. Citado 2 vezes nas páginas 19 e 32.

SILVA, T. C.; ZHAO, L. *Machine learning in complex networks*. [S.l.]: Springer, 2016. Citado na página 28.

SOUZA, A. M. d. *Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão*. Tese (Doutorado) — Mestrado em Sistemas de Informação e Gestão do Conhecimento, 2020. Citado 9 vezes nas páginas 24, 27, 28, 29, 30, 31, 32, 36 e 42.

SOUZA, R. G. D. Previsões dentro e fora da amostra da regra de taylor utilizando fatores comuns para o período de 2002: 02 à 2015: 04. In: *Anais do XLIII Encontro Nacional de Economia [Proceedings of the 43rd Brazilian Economics Meeting]*. [S.l.: s.n.], 2016. Citado na página 27.

TEIXEIRA, R. d. C. P.; MENTGES, M. J.; KAMPFF, A. J. C. Evasão no ensino superior: um estudo sistemático. *Publicação em final de outubro, 2019, Brasil.*, 2019. Citado na página 16.

TEODORO, L. de A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, v. 28, p. 838–863, 2020. Citado 2 vezes nas páginas 32 e 41.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975. Citado 6 vezes nas páginas 15, 16, 21, 23, 24 e 39.

TINTO, V. Limits of theory and practice in student attrition. *The journal of higher education*, Taylor & Francis, v. 53, n. 6, p. 687–700, 1982. Citado na página 39.

TINTO, V. *Leaving college: Rethinking the causes and cures of student attrition*. [S.l.]: University of Chicago press, 1993. Citado 2 vezes nas páginas 16 e 20.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and information systems*, Springer, v. 14, p. 1–37, 2008. Citado 2 vezes nas páginas 33 e 35.