UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

JAILMA JANUÁRIO DA SILVA

Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior

São Paulo
2022

JAILMA JANUÁRIO DA SILVA

Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Orientador: Prof. Dr. Norton Trevisan Roman

São Paulo

2022

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Silva, Jailma Januário da
Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior / Jailma Januário da Silva; orientador, Norton Trevisan Roman. -- São Paulo, 2022.
77 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Aprendizado de máquina. 2. Ensino superior. 3. Classificação multiclasse. I. Roman, Norton Trevisan, orient. II. Título.

Dedico este trabalho a meu pai (in memoriam), minha mãe e meus irmãos

Agradecimentos

Agradeço a meus pais José Januário da Silva (in memorian) e Maria Marcelino Januário, por acreditarem que a educação poderia mudar a realidade em que vivíamos, e dessa forma nos deram todo apoio e incentivo para continuar estudando, mesmo quando as circunstâncias eram difíceis.

Aos meus irmãos Josuel, Joelma, José e Joedson muito obrigada por me ajudarem nos momentos mais difíceis e por sempre me apoiarem na decisão de continuar estudando, sem vocês eu não teria conseguido chegar até aqui.

À Luzinete, uma mãe que a vida me deu, agradeço por me ajudar e sempre estar comigo em todos os momentos, e também por ser uma luz em minha vida desde o início da minha caminhada na universidade.

Agradeço a todos os meus amigos que me incentivaram e acreditaram que eu poderia conseguir, obrigada por me ouvirem nos momentos de alegria e também nos momentos mais difíceis dessa caminhada, sem vocês o caminho seria mais difícil.

Agradeço ao meu orientador professor Norton Trevisan Roman por toda a paciência e compreensão em entender minhas dificuldades e por estar sempre disposto a ajudar.

Agradeço ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Silva, Jailma Januário da. **Uma comparação de técnicas de Aprendizado de Máquina para Predição de evasão de estudantes no ensino público superior**. 2022. 76 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

A evasão de alunos dos cursos ou das instituições de ensino públicas contribui para um problema de falta de mão de obra qualificada no mercado de trabalho, pois novos profissionais deixam de ser formados e vagas que necessitam de profissionais qualificados ficam ociosas. Além disso, instituições que têm consideráveis perdas de alunos também têm perdas de verbas que poderiam ser utilizadas para mantê-las em bom funcionamento. Adicionalmente à problemática da evasão no ensino superior estão as diferentes situações em que o aluno pode estar no sistema de ensino. De acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), estas situações podem ser classificadas como: alunos com matrícula trancada, alunos desvinculados do curso (alunos evadidos) e alunos transferidos para outro curso da mesma instituição. Dessa forma, o presente trabalho tem por objetivo geral fazer a aplicação de técnicas de aprendizado de máquina em uma base de dados pública para identificar estudantes que estão em diferentes situações no ensino superior brasileiro, conforme identificadas na base de dados disponibilizada pelo INEP. Das técnicas aplicadas(árvores de decisão, Naive Bayes, Regressão Logística e Redes Neurais), as que melhor resultado apresentaram, em termos de acurácia, sensibilidade e especificidade, foram Árvores de decisão apresentando 73% de acurácia, 60% de sensibilidade e 89%de especificidade. Seguido pela técnica de regressão logística com 54% de acurácia, 55% de sensibilidade e 85% de especificidade. Por fim, foi disponibilizado o melhor modelo para a predição dos diferentes vínculos que o aluno pode ter em relação ao ensino superior.

Palavras-chaves: Aprendizado de máquina. Ensino superior. Classificação multiclasse

Abstract

Silva, Jailma Januário da. **Predição de estudantes em situações reais no ensino superior brasileiro utilizando uma base de dados pública** 2022. 76 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, DefenseYear.

The desertion of students from public educational institutions contributes to the problem of lack of qualified professionals in the laboral market, because of new professionals are no longer trained and jobs that need qualified professionals remain vacant. In addition, institutions that have considerable student losses also have lost funds that could be used to keep them working. In addition, there are different. situations in which the students may be in the education system. According to the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), these situations can be classified as: students with locked enrollment, students unlinked to a course (evaded students) and students transferred to another course of the same institution. Thus, the present work aims to apply machine learning techniques in a public database to identify students who are in different situations in Brazilian higher education, as identified in the database provided by INEP. Of the applied techniques (decision trees, Naive Bayes, Logistic Regression and Neural Networks), the ones that presented the best results, in terms of accuracy, sensitivity and specificity, were Decision Trees presenting 73% of accuracy, 60% of sensitivity and 89%specificity. Followed by the logistic regression technique with 54% accuracy, 55% sensitivity and 85% specificity. Finally, the best model was made available for the prediction of the different bonds that the student may have in relation to higher education.

Keywords: Machine learning. Higher Education. Multiclass classification

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

KDD	Knowledge Discovery in Databases
EDM	Educational Data Mining
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ABRES	Associação Brasileira de Estágios

Sumário

1 Introdução

Diariamente uma grande quantidade de dados é produzida em diferentes setores, como por exemplo na saúde, nas indústrias e também na educação. Nesses dados estão contidas informações fundamentais, que vão desde quem atua ativamente na área até quem é consumidor passivo dos serviços e/ou produtos oferecidos. Tais informações precisam ser analisadas de modo a ser descoberto conhecimento útil para as organizações e, dessa forma, possam melhorar seus serviços e produtos, bem como subsidiar as tomadas de decisões nestes setores (COSTA *et al.*, 2013).

Diante deste cenário, a área de mineração de dados teve um notável avanço desde seu surgimento. A mineração de dados faz parte de um processo de descoberta de conhecimento em uma base de dados, denominado KDD (Knowledge Discovery in Databases), que consiste em seis etapas: a primeira etapa do processo de KDD consiste na definição dos objetivos que se deseja alcançar com a aplicação do processo; a segunda etapa consiste em selecionar os dados; durante a terceira etapa ocorre o pre-processamento (limpeza) dos dados; a quarta etapa consiste em transformar os dados em um padrão que possa ser utilizado nos algoritmos aplicados; durante a quinta etapa, já com os dados devidamente preparados, é realizada a mineração de dados,que consiste em aplicar técnicas de aprendizado de máquina como classificação, regressão ou clusterização, a técnica é escolhida de acordo com o objetivo, e a partir dessa aplicação os padrões são descobertos; a sexta e última etapa consiste em avaliar e interpretar os resultados (e.g. (PAZ; CAZELLA, 2017; SILVA; PINTO *et al.*, 2018)).

De acordo com Ahmed e Khan (2019), ao observar as particularidades dos dados vindos de contextos educacionais, pesquisadores da área de informática na educação têm se dedicado a aplicar técnicas de mineração de dados para realizar a análise desses dados de forma adequada. Com isso surge a linha de pesquisa denominada mineração de dados educacionais, a qual tem por objetivo "desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais" (COSTA *et al.*, 2013). Ter acesso a uma análise adequada de dados vindos desse contexto pode fazer com que a gestão da instituição consiga desenvolver ações que possam melhorar tanto a aprendizagem do aluno, como evitar que o mesmo

venha a evadir das universidades, na modalidade presencial e a distância (e.g. (COSTA *et al.*, 2013; BAKER; ISOTANI; CARVALHO, 2011)).

Diversos países, como Peru, Equador e Colômbia, têm utilizado a mineração de dados educacionais para ajudar na tomada de decisão das instituições em problemas como a evasão escolar, tanto na educação básica como na educação superior (e.g. (ROCHA *et al.*, 2017; MAYRA; MAURICIO, 2018; PEREIRA; ZAMBRANO, 2017)). O Brasil também está entre os países que utilizam desses recursos computacionais para auxiliar na tomada de decisão na educação de forma mais efetiva como demostrado em (e.g. Lanes e Alcântara (2018), Silveira R.F. Victorino (2019)). Com isso, diversos pesquisadores em diferentes universidades analisam os dados no contexto em que a instituição está inserida, com diferentes objetivos, que vão desde contribuir com o ensino e aprendizagem do sistema de ensino, até predizer quais alunos estão propensos a evasão (BAKER; ISOTANI; CARVALHO, 2011).

1.1 Motivação e justificativa

Até o ano de 2018, o número de ingressantes no ensino superior era de 3.445.935, com apenas 36,68% conseguindo concluir o curso e obter seu diploma. Estes são dados divulgados pela Associação Brasileira de Estágios (ABRES)[1], e que demonstram um triste cenário para um país que está em desenvolvimento, uma vez que a educação tem ligação direta com o avanço de um país. Dessa forma, é preciso entender o que está levando tantos brasileiros a desistirem do ensino superior, para que possam ser realizadas ações que consigam reduzir este problema.

A evasão de alunos, seja do curso ou da instituição de ensino (privada ou pública), contribui com o problema de falta de mão de obra qualificada no mercado de trabalho, pois novos profissionais deixam de ser formados e vagas que necessitam de profissionais qualificados ficam ociosas. Além disso, instituições que têm consideráveis perdas de alunos também têm perdas de verbas que poderiam ser utilizadas para mantê-las em bom funcionamento (PRESTES; FIALHO, 2018).

Pensando no atual cenário do ensino superior brasileiro, e em todas as consequências que perder um aluno traz às instituições e à sociedade, diferentes universidades começaram a buscar a causa da desistência de seus alunos. Com isso, foram realizados alguns estudos

[1] ⟨https://abres.org.br/⟩

(e.g. (GONÇALVEZ T.C. SILVA, 2018; SILVEIRA R.F. VICTORINO, 2019; DIGIAMPIETRI; NAKANO; LAURETTO, 2016)) sobre a evasão no ensino superior brasileiro, estudos que vão desde descobrir os fatores que levam o aluno a desistir, até a utilizar soluções computacionais avançadas para este problema. Como temos uma diversidade regional no país, cada instituição leva em conta as particularidades da região em que está inserida, sendo dessa forma difícil estabelecer um padrão comum de análise e solução a esta problemática.

Além disso, O aluno pode estar em diferentes categorias conforme o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), como vínculos finais que também contribuem para a identificação da evasão. São elas: alunos com matrícula trancada, desvinculados do curso e transferidos para outro curso da mesma instituição. Dessa forma, justifica-se esta pesquisa, cujo objetivo foi realizar uma classificação multiclasse e verificar se é possível predizer a situação futura de cada aluno no ensino superior.

1.2 Problema de pesquisa

Atualmente, estudos que tratam da situação do aluno no ensino superior focam em fazer a predição de evasão a partir de duas classes: evadidos e não evadidos (e.g. (MAYRA; MAURICIO, 2018; PEREZ; CASTELLANOS; CORREAL, 2018; SILVEIRA R.F. VICTORINO, 2019)). No entanto, os dados disponíveis no site do INEP, sobre o ensino superior brasileiro, não classificam a situação dos alunos no curso em apenas duas categorias. Conforme o INEP, um aluno pode estar com a matrícula trancada, desvinculado do curso, transferido para outro curso da mesma instituição ou tê-lo concluído.

Dessa forma, identificar antecipadamente grupos de alunos que pertençam a tais situações também traz benefícios à instituição, e pode fazer com que a gestão desenvolva ações de forma mais assertiva que permitam que o aluno continue o ensino e chegue à conclusão do seu curso.

Diante do exposto, elaborou-se a seguinte questão de pesquisa, que este trabalho buscou responder: É possível classificar alunos com vínculos reais no sistema de ensino superior, utilizando os dados disponíveis pelo INEP?

1.3 Objetivo

- **Objetivo Geral**:

  Aplicar técnicas de aprendizado de máquina em uma base de dados pública para classificar situação de estudantes no ensino superior brasileiro.

- **Objetivos Específicos**:

  Verificar se é possível classificar o aluno em diferentes categorias usando os dados de instituições públicas disponibilizados pelo INEP;

  Aplicar as técnicas mais comuns em aprendizado de máquina para evasão de estudantes do curso, ou instituição;

  Escolher o melhor modelo para a categoria de problemas multiclasse, com base nas métricas selecionadas.

1.4 Organização do documento

Este documento está organizado da seguinte forma. Na Seção 2 é apresentada uma revisão sistemática da literatura, realizada com o intuito de identificar quais as técnicas mais utilizadas para resolver esta problemática. Na Seção 3 são apresentados os conceitos que são fundamentais para o entendimento da pesquisa. Na Seção 4 é descrita detalhadamente a metodologia a ser seguida. Na Seção 5 são apresentados os resultados obtidos com o estudo. Por fim na Seção 6 é apresentada a conclusão e trabalhos futuros.

2 Revisão da literatura

Nesta seção, é apresentada uma revisão sistemática da literatura sobre predição de estudantes com risco de evasão no ensino superior, realizada com o objetivo de destacar pesquisas atuais relacionadas ao uso de aprendizado de máquina e técnicas de mineração de dados para identificar estudantes que tenham risco de abandonar o ensino superior. A revisão sistemática da literatura é definida como uma metodologia de pesquisa específica, desenvolvida para reunir e avaliar as evidências disponíveis relacionadas a um tópico (BIOLCHINI J.; TRAVASSOS, 2005). De acordo com Kitchenham (2004), as revisões sistemáticas são úteis para a identificação, avaliação e interpretação das pesquisas disponíveis, específica para algum tópico, questão de pesquisa ou fenômeno de interesse .

2.1 Protocolo da revisão sistemática

O protocolo utilizado nesta pesquisa é baseado no modelo apresentado por Biolchini J. e Travassos (2005), Kitchenham (2004), onde é necessário estabelecer o objetivo da pesquisa, as questões a serem respondidas nos trabalhos encontrados, definição dos critérios de inclusão, exclusão e qualidade, bem como definir os bancos de dados a serem pesquisados, juntamente com as strings de busca e as palavras-chave (previamente determinadas a partir dos resultados de uma revisão exploratória).

Objetivo da revisão sistemática

Esta revisão sistemática teve como objetivo identificar os diferentes métodos e técnicas de mineração de dados e aprendizado de máquina usadas para prever o risco de evasão dos alunos nas instituições de ensino superior em cursos na modalidade presencial.

Questões de pesquisa:

Questão1: Quais informações presentes nas bases de dados são utilizadas para serem analisadas?

Esta questão tem como objetivo identificar quais são os principais dados utilizados para treinar os modelos e obter os melhores desempenhos na predição de evasão do aluno, uma vez que as instituições possuem diferentes informações, desde socioeconômicas ao histórico escolar do aluno.

Questão2: Quais abordagens ou técnicas são as mais utilizadas?

Nesta questão buscou-se entender quais técnicas são as mais aplicadas a este tipo de problema, e quais demostram os melhores resultados. Também será possível entender o que faz com que tais técnicas tenham os melhores resultados com relação a dados educacionais.

Questão3. Quais métodos ou técnicas são usados para avaliar as abordagens adotadas?

Nesta questão buscamos identificar quais são as métricas mais utilizadas para avaliar o modelos desenvolvidos, considerando as particularidades de dados educacionais, e por que essas métricas são as mais empregadas.

2.2 Critérios de inclusão, exclusão e qualidade

Foram definidos os seguintes critérios de inclusão, exclusão e qualidade:

- Critérios de inclusão:
  1. Artigos revisados por pares, publicados a partir de 2015 até o ano de 2020.
  2. Estudos descrevendo métodos e/ou técnicas destinadas a prever os alunos com risco de evasão no ensino superior.
- Critérios de exclusão:
  1. Estudos não relacionados ao ensino superior;
  2. Estudos que abordem a modalidade a distância;
  3. Artigos curtos (menos de 4 páginas);
  4. Estudos repetidos, caso em que somente a primeira fonte de pesquisa será considerada
- Critérios de qualidade:
  1. O estudo apresenta algum método, técnica ou ferramenta para avaliar seus resultados e abordagens?
  2. O estudo tem um objetivo de pesquisa bem definido e/ou perguntas baseadas na literatura relacionada?
  3. O estudo compara seus resultados com os da literatura relacionada?
  4. O estudo foi realizado em um ambiente de gestão educacional?

Os critérios de qualidade possuem uma pontuação associada, de modo a melhor filtrar os estudos, após ocorrer a leitura do mesmo. Dessa forma, se um dos estudos não

atingir a pontuação mínima de 1 ponto nos critérios de qualidade, este será excluído. Na tabela 1 a seguir é apresentada a pontuação para cada critério de qualidade.

Tabela 1 – Critério de qualidade e a pontuação definida

Critério	***Pontuação***
1	0,5
2	1
3	0,5
4	0,5

Fonte – Jailma Januário, 2021

2.3 Repositórios pesquisados

A busca foi realizada nas bases de dados Scopus, por ser um banco de dados que possui artigos, jornais e revistas na área científica, técnica, etc, bem como, a base de dados IEEE Xplore Digital Library, também por ser um banco de dados que integra artigos de periódicos, anais de conferências relacionados a ciência da computação, engenharia elétrica e eletrônica. Foi elaborada uma *string* de busca, a qual teve algumas alterações entre cada base de dados, a partir de uma pesquisa exploratória, onde foi possível identificar os termos mais usados em trabalhos com a mesma temática. A tabela 2, a seguir, apresenta as bases de dados com a string utilizada. Foram considerados trabalhos escritos em inglês, porque é o idioma mais presente nas bases utilizadas.

A busca na base de dados IEEE Xplore Digital Library retornou 1.147 artigos. Após a filtragem por período, idioma, e palavras chaves no título dos trabalhos, foram obtidos 575 artigos. Após a aplicação dos critérios de inclusão e exclusão, restaram 10 artigos, cujo conteúdo foi lido na íntegra. A busca na base de dados SCOPUS retornou 128 artigos. Após a filtragem por período, idioma e palavras chaves, foram obtidos um total de 57 artigos, que foram analisados de acordo com os critérios de exclusão e inclusão, restaram 11 artigos para serem lidos na íntegra. Após os estudos passarem pelo filtro e inclusão e exlusão, foram aplicados os critérios de qualidade e suas respectivas pontuações.

Tabela 2 – Bases de dados e suas string de busca utilizadas

Bases de dados	***strings utilizadas***
SCOPUS	(TITLE-ABS-KEY ("Data mining"OR "educational data mining"OR "machine learning"OR "prediction techniques") AND TITLE-ABS-KEY ("university"OR "graduate") AND TITLE-ABS-KEY ("dropout"OR "evasion"))
IEEE Xplore Digital Library	((("All Metadata":"Data mining"OR "educational data mining"OR "machine learning"OR "prediction techniques") AND "All Metadata":"university"OR "graduate") AND "All Metadata":"dropout" OR "evasion")))

Fonte – Jailma Januário, 2021

2.4 Resultados

Nesta seção são apresentados os resultados da revisão sistemática, dos 21 artigos que foram selecionados. Os resultados serão apresentados de acordo com cada questão de pesquisa.

Questão 1: Que tipos de dados são usados para treinar e testar modelos?

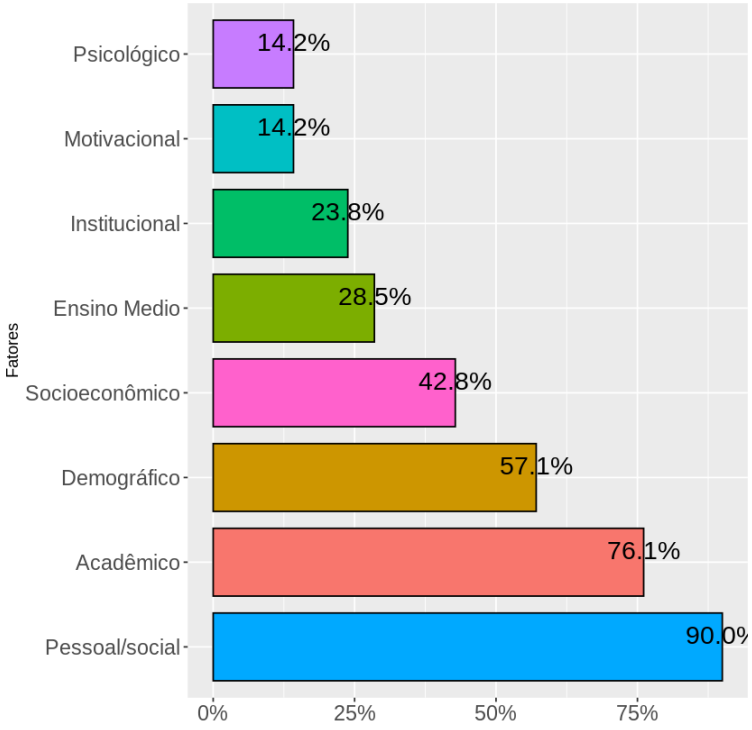
De acordo com os estudos pode-se, identificar vários fatores que podem ser considerados contribuintes diretos para a evasão, seja do curso ou da instituição, tais como:

- Fatores motivacionais: referem-se à causa da escolha inicial do aluno pela instituição e/ou curso.
- Fatores psicológicos: referem-se a questões de autonomia e resiliência;
- Fatores institucionais: referem-se a dados da universidade, como número de cursos, departamentos, campi, etc.
- Características acadêmicas: referem-se às notas nos cursos, número de cursos concluídos, quantidade de reprovações, histórico acadêmico, etc.
- Fatores socioeconômicos: referem-se à renda familiar, assistência ao aluno, se o aluno trabalha ou não;
- Fatores demográficos: referem-se à distância da casa do aluno à universidade, se o aluno veio de outra cidade, estado ou país;
- Fatores pessoais/sociais: referem-se às informações pessoais do aluno, se ele mora sozinho ou divide uma casa com outros alunos, seu estado civil. etc.

- Anterior ao ingresso no curso: refere-se a questões relacionadas ao tipo de ensino médio que o aluno frequentou antes de entrar no ensino superior, juntamente com o método pelo qual ele ingressou no ensino superior.

O número de artigos, agrupados por característica estudada, é mostrado na Figura 1. Como se constata, Características Pessoais/Sociais (apresentadas em 19 artigos - 90,4%), e Características Acadêmicas (apresentadas em 16 artigos - 76,1%), são as Características mais analisadas na pesquisa para identificação de grupos de risco de evasão, seguidas por Características Demográficas e Socioeconômicas (12 (57,1%) e 9 (42,8%) artigos, respectivamente). No outro extremo, os fatores Psicológicos e Motivacionais são os recursos menos analisados, ambos compreendendo 3 (9,5%) artigos cada.

Figura 1 – Porcentagem de artigos por fatores



Fonte – Jailma Januário, 2021

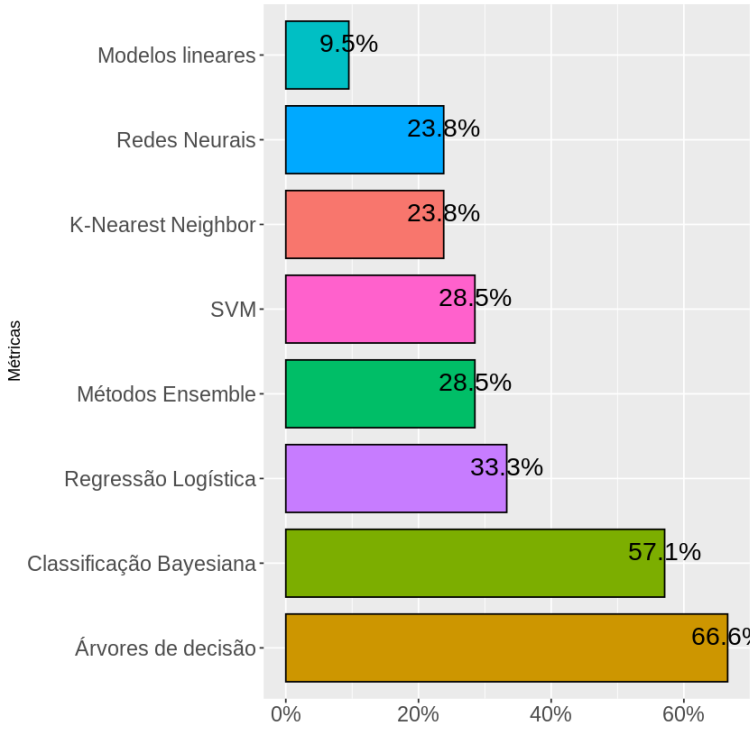
Os dados usados nas análises foram obtidos tanto de ambientes de gestão educacional, ou seja em sistemas da própria instituição, quando por meio de questionário enviados a ex alunos (menos comum nesse tipo de estudo), ou em algum repositório disponibilizado pelo governo. A quantidade de atributos em um conjunto de dados varia, mas a alta dimensionalidade predomina neste tipo problema. Diversas técnicas são utilizadas para reduzir o conjunto de dados, porém em alguns trabalhos como o realizado por Perez, Castellanos e

Correal (2018), quando são comparados os modelos construído com conjunto de dados completos e conjunto de dados utilizando alguma técnica de redução de dimensionalidade, os melhores modelos são os que contém conjunto de dados completos.

Questão 2: Quais abordagens ou técnicas são utilizadas com mais frequência?

Com esta questão, tentamos determinar quais técnicas de aprendizado de máquina são mais populares entre os pesquisadores, a fim de identificar boas técnicas para serem reproduzidas neste trabalho. A maioria dos artigos analisados implementa mais de três técnicas de aprendizado de máquina, de forma a poder comparar os modelos entre si. A Figura 2 mostra a fração de artigos que implementam cada uma das técnicas identificadas.

Figura 2 – Porcentagem de artigos por técnica utilizada



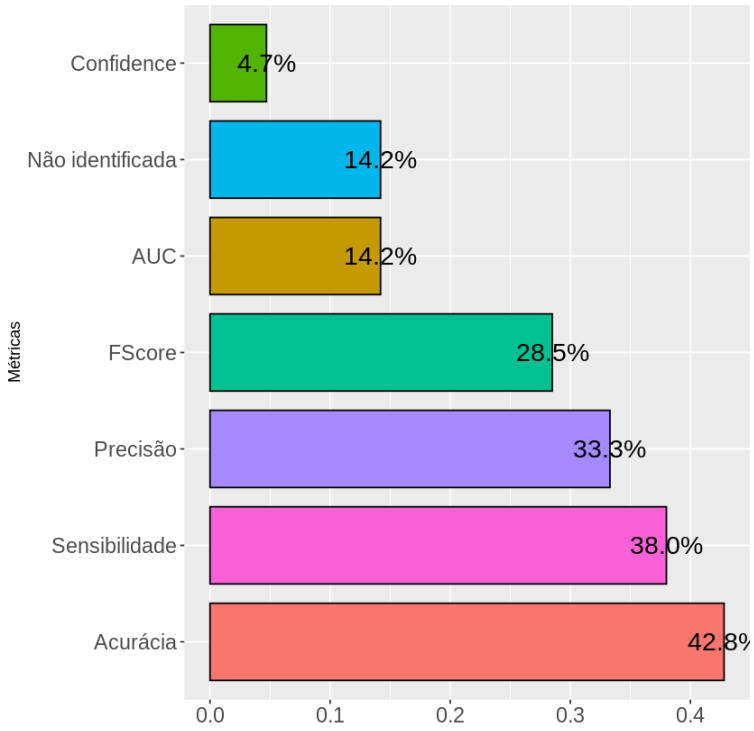
Fonte – Jailma Januário, 2021

Conforme mostrado na figura, as árvores de decisão se destacam como a abordagem mais popular, estando presente em dois terços dos artigos (66,6%, correspondendo a 14 artigos no total). Na sequência, encontram-se métodos de classificação bayesiana (mais especificamente, Naïve Bayes e Redes Bayesianas), aparecendo em 12 artigos (57,1% do total), Regressão Logística (33,3% - 8 artigos), SVM e Modelos essembles (28,5% cada - 6 artigos) e, por fim, os modelos de Vizinho Mais Próximo e Redes Neurais (tanto Perceptron Multicamadas quanto Redes Neurais Profundas), com 5 artigos cada (23,8% do total) e Modelos Lineares, com 2 artigos (9,5%).

Questão 3: Quais métodos ou técnicas são usados para avaliar as abordagens adotadas?

Com esta questão, buscou-se compreender como os modelos gerados são avaliados e quais recursos de avaliação são levados em consideração para tal. Conforme mostrado na Figura 2, a acurácia destaca-se como a métrica de avaliação mais relatada, sendo usada em 42,8% dos artigos (9 artigos no total). Na sequência encontra-se a sensibilidade, com 38% (8 artigos), precisão, com 33,3% (7 artigos), e F1-score, utilizada em 28,5% dos artigos (6 artigos no total).

Figura 3 – Porcentagem de artigos por métrica utilizada



Fonte – Jailma Januário, 2021

No outro extremo, a área sob a curva ROC (AUC) foi utilizada em 3 artigos (14,2%), com um único artigo (4,7%) apresentando intervalos de confiança para os valores relatados. Embora minoria em nossa pesquisa, 14,2% dos artigos (3 no total) não apresentaram medida de avaliação de seus resultados. Por fim, como mostra a Tabela 3, percebe-se que os artigos analisados relatam majoritariamente mais de uma métrica de avaliação. Mesmo que um terço deles prefira se ater a uma única métrica, quase a mesma quantidade faz uso de mais de três diferentes métricas para avaliar os resultados.

Tabela 3 – Número de artigos por métrica utilizada

Artigo	***Número de métricas***
3	0
7	1
4	2
1	3
6	> 3

Fonte – Jailma Januário, 2021

2.5 Outros trabalhos relacionados

O trabalho realizado por Ahmed e Khan (2019) em Bangladesh, país que, assim como o Brasil e outros países, sofre com altas taxas de evasão no ensino superior, buscou predizer alunos com risco de evasão usando técnicas de mineração de dados educacionais (EDM) no País. O autor utilizou dados de 480 estudantes de diferentes cidades de Bangladesh, que foram adquiridos através de um questionário enviado a ex-alunos. O questionário continha 28 questões, que abordavam tanto fatores pessoais, acadêmicos como institucionais.

Os dados foram divididos em três conjunto de dados, um com dados pessoais, outro com dados acadêmicos e outro com todos os tipos de dados (acadêmicos, pessoais e institucionais), para dessa forma escolher o melhor modelo. O autor utilizou cinco métodos de aprendizado de máquina: vizinhos mais próximos, árvores de decisão, métodos de ensemble, redes neurais, classificação bayesiana e regressão logística. Para avaliar os modelos foram utilizadas as métricas de acurácia, sensitividade, precisão e f-score. Dentre os modelos, o que teve melhor desempenho foram as redes neurais, com acurácia de 0.915 e demais métricas com valores parecidos, para o conjunto de dados que continha todos os dados que abordavam os diferentes fatores.

O estudo feito por Nagy e Molontay (2018), utilizou dados da escola secundária (ensino médio) dos alunos de uma universidade em Budapeste, para prever os alunos que tendem a evadir no ensino superior. O autor justifica a escolha dessa abordagem pelo fato de que na Hungria os alunos que não concluem os cursos desistem no primeiro ano. Com isso, fazer essa previsão o quanto antes, com base em dados anteriores ao ingresso no curso superior, pode evitar que essa situação aconteça.

Foram utilizados dados do ensino médio e pessoais de 15.825 estudantes em um período de 7 anos (2010 a 2017), fornecidos pela Universidade de Tecnologia e Economia de Budapeste. O conjunto de dados estava composto por 40 atributos, apresentando uma alta dimensionalidade, característica típica desse tipo de problema. Para reduzir a dimensionalidade da base de dados, o autor aplicou a técnica de Análise de Componentes Principais (PCA), que resultou em apenas 7 componentes principais que podem explicar a variação do conjunto de dados. Com isso o autor utilizou dois conjunto de dados: o completo, em que no pré-processamento foram imputados valores ausentes, e outro com 7 componentes que resultou da aplicação da técnica de PCA.

Para cada conjunto de dados as seguintes técnicas de aprendizado de máquina foram utilizadas: árvores de decisão, redes bayesianas, modelos lineares, classificação bayesiana, métodos ensemble, k-NN, regressão logística e aprendizado profundo, este último com melhor performance entre os modelos, com 81% de acurácia e 70,8% de precisão no conjunto de dados completo, já no conjunto de dados com as características filtradas este mesmo modelo obteve um mau desempenho. Para este conjunto de dados o modelo que obteve uma melhor sensibilidade foi a árvore de decisão, com 89,7%. Para avaliar os modelos, foram utilizadas as métricas convencionais para esse tipo de problema: precisão, sensibilidade, acurácia e AUC.

Na sequência o artigo de Perez, Castellanos e Correal (2018) relata um estudo realizado em uma universidade estadual da Colômbia, com dados coletados do sistema acadêmico de 2004 a 2010. o conjunto de dados contém 31 variáveis, abordando fatores pessoais, acadêmicos e socioeconômicos. Foram utilizados dois conjunto de dados, um com 31 variáveis, e um com a aplicação da técnica de PCA, que reduziu o conjunto de dados em 15 componentes principais que explicam 93,21% da variância dos dados.

As técnicas utilizadas foram árvores de decisão, classificação bayesiana e regressão logística. Para a comparação dos modelos o autor utilizou a Métrica *Receveier Operating Characteristic* (ROC). Os modelos com os dados originais sem o PCA tiveram um melhor desempenho, em particular a árvore de decisão com ROC- AUC de 0,94%. O autor considera essa diferença de desempenho entre o conjunto de dados original e com o PCA, ser devido o conjunto de dados com PCA apresentar uma perda significativa da variância para os 15 PCs encontrado.

O estudo realizado por Rocha *et al.* (2017), por sua vez ressalta os malefícios desta problemática também vivenciada no Peru, principalmente a perda de investimentos públicos.

Dessa forma, o autor realiza um estudo de caso na universidade internacional de San Marcos, no curso de engenharia de sistemas. O autor utiliza três técnicas de aprendizado de máquina, classificação bayesiana, árvores de decisão e redes neurais artificiais com multilayer perceptron, além de uma abordagem híbrida.

Os dados utilizados foram disponibilizados pela referida universidade e contêm um total de 1.154 registros, com 89 fatores que abordam: fatores pessoais/sociais e acadêmicos. Após uma filtragem, eliminando anomalias e variáveis com muitos valores ausentes, o autor utilizou um conjunto de dados com 24 características. A métrica de avaliação dos modelos que o autor utiliza para encontrar o melhor modelo é a métrica de acurácia, e o modelo que apresentou a maior acurácia foi a técnica de redes bayesianas.

O estudo de Rodríguez-Muñiz *et al.* (2019) teve por objetivo fazer uma previsão de alunos em risco de evasão em uma universidade espanhola (universidade de Oviedo), utilizando métodos como árvores de decisão, redes bayesianas e support vector machine (SVM) e métodos ensembles. O conjunto de dados continha 40 atributos, e foi disponibilizado pela referida universidade, sendo também composto por preenchimento de um questionário enviado aos alunos, abordando fatores acadêmicos, pessoais, socioeconômicos e motivacionais.

Cada algoritmo utilizado para a implantação das técnicas foi treinado usando suas configurações padrão. Para a avaliação dos modelos, foram utilizadas as métricas de *f-meansure e accuracy*. Para este conjunto de dados, a técnica que teve uma maior acurácia foi a de métodos ensemble, com 86,6% de acurácia. Além disso, o autor apresenta um novo fator que contribui para a evasão em relação a literatura atual, como saber se o aluno estuda em tempo integral ou parcial, e também o fato do estudante estar em um curso de meio período, e a razão pela escolha da universidade.

O trabalho de Sultana, Khan e Abbas (2017) usou características cognitivas (dados sociais, acadêmicos, socioeconômicos) e não cognitivas (comportamento e atitude do aluno em um ambiente, gerenciamento de tempo, auto avaliação, etc.) para construir o modelo de previsão de evasão. As técnicas utilizadas foram árvores de decisão, regressão logística, classificação bayesiana e redes neurais.

Os dados foram adquiridos do curso de engenharia elétrica da universidade nacional de ciências e tecnologia no Paquistão. Estes dados formaram o conjunto de dados, que possuía 113 instâncias, abordando fatores pessoais, acadêmicos, socioeconômicos. Os dados não cognitivos foram coletados de uma amostra de participantes-alvo.

A métrica apresentada para avaliar os modelos foi a medida de acurácia. A técnica que teve maior acurácia foi árvore de decisão, apresentando 61% com a previsão apenas com dados cognitivos e 65% quando os dois tipos de dados (cognitivos e não cognitivos) agrupados, mais especificamente atributos que envolviam questões de autoavaliação ou liderança.

O estudo de caso realizado por Hutagaol e Suharjito (2019) utilizou técnicas como vizinhos mais próximos, redes bayesianas, árvores de decisão e métodos ensembles para predizer a evasão de estudantes em uma universidade da indonésia, fazendo uma comparação entre as técnicas para escolher o melhor modelo para este problema. Os dados foram adquiridos do sistema da própria universidade, e continham 17 atributos que abordaram fatores pessoais, demográficos e acadêmicos. Na fase de processamento dos dados, o autor utilizou a técnica de Learning Vector Quantization, para selecionar os atributos mais importantes com pontuação maior que 50%. Com isso, o conjunto de dados final é composto por 13 variáveis, relativamente pequeno se comparado com os demais estudos descritos.

Para avaliar os modelos, as técnicas aplicadas foram as medidas convencionais para este tipo de problemas, sendo elas acurácia, precisão, sensitivity e specitivity. Para este estudo, o método ensemble foi o melhor modelo, com 0,9882 de acurácia e recall de 0,9290, seguido da classificação bayesiana, com 0,9824 de acurácia, e 0,9360 de recall, demonstrando que o uso de um meta-classificador é melhor que um classificador único.

O trabalho de Mayra e Mauricio (2018) é um estudo de caso realizado no Equador, que analizou os dados de 3.000 estudantes universitários da Universidade de Cotopaxi. A autora definiu 11 fatores com base na literatura relacionada. O conjunto de dados foi composto por 17 variáveis e as técnicas utilizadas foram regressão logística, árvores de decisão e máquina de vetores de suporte (SVM). A avaliação apresentada dos modelos foi a porcentagem de acurácia. Dentre os modelos, o classificador de árvore de decisão teve a maior acurácia apresentada (98%). A autora considera a técnica de árvore de decisão ideal para este tipo de problema, devido a fácil interpretabilidade dos resultados.

Por fim o trabalho de Silveira R.F. Victorino (2019) também realizou um estudo de caso no Brasil, na Universidade de Brasilia (UNB). Nele, o autor utilizou técnicas de mineração de dados para analisar o perfil alunos do curso de engenharia, de 2009 a 2019, além de apresentar a metodologia CRISP-DM (Cross Industry Process Model for Data Mining), que foi seguida para a realização do trabalho. Foram coletados dados

pessoais/sociais, demográficos e de desempenho acadêmico, além de alguns dados anteriores ao ingresso no curso, a partir do sistema da própria instituição. O autor utilizou a técnica de ensembles e de modelos lineares, com três algoritmos diferentes: Modelo Linear Generalizado (GLM), algoritmo de Boosting (GBM) e Random Forest (RF), todos com suas configurações padrão. O melhor modelo para este estudo de caso foi o modelo linear generalizado, com 86,56% de acurácia, única métrica apresentada no artigo.

2.6 Considerações finais

O problema da evasão de alunos em instituições de ensino superior no Brasil e em outros países vem sendo discutido em larga escala, como pode-se perceber nesta revisão sistemática da literatura. Diversas técnicas são aplicadas para solucionar este problema, desde as mais simples às mais avançadas. No entanto, a mais comum entre os autores é a técnica de árvore de decisão, A razão para isto está na facilidade de interpretação dos resultados, fácil implementação e a forma de lidar bem com valores faltantes. Outras técnicas, como ensembles e redes neurais, quando utilizadas superam as técnicas de árvore de decisão, porém são menos populares entre os autores para este tipo de problema.

As métricas utilizadas para avaliar os modelos são acurácia, sensibilidade, precision e f score, métricas comuns em problemas de aprendizado de máquina. Assis (2018) destaca que a métrica de sensibilidade/recall é de extrema importância em estudos desse tipo, pois deve-se evitar o máximo possível de falsos negativos, ou seja alunos que estão em risco de evasão sendo classificados erroneamente.

Os fatores pessoais/sociais e acadêmicos são os mais utilizados para compor os vetores de características utilizados para treinar os modelos. Apesar dessa preferência, outros fatores também são explorados, como demográficos e socioeconômicos, entre outros. O número de atributos também varia consideravelmente, de 11 a 44, embora alguns estudos apliquem técnicas de redução de dimensão para diminuir esse valor.

Estes trabalhos relaciona-se com esta pesquisa, uma vez que trata-se da mesma problemática. Porém a maioria desses estudos são realizados dentro de uma instituição, mesmo quando os dados são adquiridos de forma pública agrupadas de diferentes instituições de um país. As razões que são consideradas para isto é que cada instituição tem suas particularidades, que devem ser consideradas quando trata-se deste tipo de problema.

Os poucos estudos que analisam os dados de todo o país para construir seus modelos obtêm ótimos resultados, como o estudo descrito por Ahmed e Khan (2019). No entanto o autor não utilizou apenas dados encontrados nas bases de dados disponíveis no país, mas também enviou um questionário a alguns alunos selecionados para contribuir com a pesquisa.

Dessa forma, este trabalho se propõe a fazer esta análise de dados educacionais, utilizando apenas a base de dados pública disponibilizada pelo governo, aplicando as técnicas de análise de dados mais utilizadas na literatura, a fim de identificar grupos de alunos que estejam em diferentes vínculos no sistema de ensino brasileiro.

3 Fundamentação teórica

Nesta seção serão apresentados os principais conceitos que são necessários para o entendimento do trabalho. São apresentados os conceitos de evasão, aprendizado de máquina, classificação, classificação multiclasse, as técnicas utilizadas: árvore de decisão, naive bayes, redes neurais e regressão logística, e as principais métricas para avaliação dos modelos.

3.1 Evasão

Segundo a Comissão Especial de Evasão do Ensino Superior (CEEES), [1] a evasão se caracteriza pela saída do aluno de seu curso de origem, sem concluí-lo. Considerando a evasão de curso que é abordada neste trabalho, existem outros níveis de evasão no sistema de ensino como a evasão da instituição e a evasão do sistema.

- Evasão da instituição: Se refere à saída do aluno do curso e de uma instituição para outra;
- Evasão do sistema: Se refere à saída do aluno de forma voluntária ou involuntária do sistema de ensino. Nesse tipo de evasão é considerado que o aluno desistiu de estudar, saindo de vez ou temporariamente do ensino superior (SOUZA, 2020);
- Evasão de curso: se refere a saída do aluno de seu curso sem finalizá-lo.

Em um estudo preliminar sobre a evasão, realizado em 1995, a CEEES utilizou uma técnica para a identificação da evasão de curso chamada técnica de acompanhamento de estudantes, que pode ser expressada pela seguinte equação:

$$\%Evasão = \frac{(N_i - N_d - N_r) * 100}{N_i}$$

Onde, N_i é a quantidade de alunos no curso, N_d é a quantidade de alunos diplomados, N_r é a quantidade de alunos retidos (atrasados no curso). Dessa forma, a evasão é calculada pela razão entre o total de alunos ingressantes (N_i) menos o total de alunos diplomados (N_d) menos o total de alunos retidos (N_r) e o total de alunos ingressantes (N_i), multiplicado

[1] ⟨http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=24676⟩

por 100. Nesse modelo para o levantamento da evasão de curso se considera a turma de alunos ingressantes e o tempo máximo do curso.

Tendo em vista que a jornada do aluno no ensino superior é complexa, esta categoria de evasão pode ter diferentes causas como "abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso) e exclusão por norma institucional" (CEEES). Essas causas qualificam uma evasão permanente do aluno com relação ao curso, outra forma de evasão existente é a evasão temporária, caracterizada no sistema pelo trancamento do curso, ou seja, quando o aluno realiza a ação de trancar o curso, ele pode ou não voltar para o mesmo. Esse conceito de temporaneidade foi demostrado no estudo realizado por Silva, Cabral e Pacheco (2016), que apresenta além das dimensões de evasão de curso, instituição e sistema, um conceito de evasão em relação ao tempo, sendo elas definitiva ou temporária. O quadro a seguir apresenta um resumo das dimensões e temporaneidade da evasão no sistema de ensino.

Figura 4 – Evasão por tempo e dimensão

	Evasão em relação ao tempo		Evasão em relação às dimensões		
	Definitiva	Temporária	Evasão do Curso	Evasão da Instituição	Evasão do Sistema de Ensino
Transferência para outro curso da mesma instituição	✓		✓		
Transferência para curso de outra instituição	✓		✓	✓	
Abandono	✓		✓	✓	✓
Desistência formal	✓		✓	✓	✓
Cancelamento de matrícula por iniciativa da instituição	✓		✓	✓	✓
Jubilamento	✓		✓	✓	✓
Trancamento		✓	✓	✓	✓

Fonte – Silva et al (2016)

Observa-se que a evasão de curso ocorre em diferentes categorias, onde apenas o trancamento é classificado como uma evasão temporária, que ocorre quando o aluno cancela a matricula em um determinado período de tempo. Nesse caso, o aluno pode ou não voltar ao curso. Outro ponto importante é que, se o aluno evadir do sistema de ensino, automaticamente ocorre a evasão da instituição e do curso. No entanto se o aluno evade do curso, nem sempre ele vai evadir da instituição e/ou do sistema de ensino (SILVA; CABRAL; PACHECO, 2016).

3.2 Aprendizado de máquina

Aprendizado de máquina pode ser definido como uma área da ciência da computação cujo objetivo principal é a criação de sistemas que tenham a capacidade de aprender com dados (SOUZA, 2020). Para Mitchell (1997), o aprendizado de máquina empenha-se em criar programas na computação que melhoram automaticamente, com dados que contenham exemplos de experiências passadas, para a solução de um determinado problema.

Com o sucesso de aplicações na área, os benefícios do aprendizado de máquina foram para além do campo da ciência da computação, tendo aplicações em diversos campos como a educação, através da construção de modelos preditivos para detecção de alunos com risco de evasão e o desenvolvimento de aplicações para melhoria no processo de ensino e aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

A área de aprendizado de máquina possui diferentes técnicas que podem ser agrupadas em três categorias, podendo variar de acordo com a cada do autor. Neste trabalho serão consideradas as categorias de acordo com Souza (2020), sendo elas: aprendizado supervisionado, onde as categorias que uma nova instância pode ser classificada são conhecidas no conjunto de dados; Aprendizado não supervisionado, em que as categorias de uma classe não são conhecidas no conjunto de dados; E aprendizado Semi-supervisionado, onde o aprendizado ocorre com características do aprendizado supervisionado e não supervisionado. Para este trabalho serão utilizadas técnicas de aprendizado supervisionado, pois as categorias em que um aluno pode ser classificado são conhecidas na base de dados.

3.3 Aprendizado supervisionado

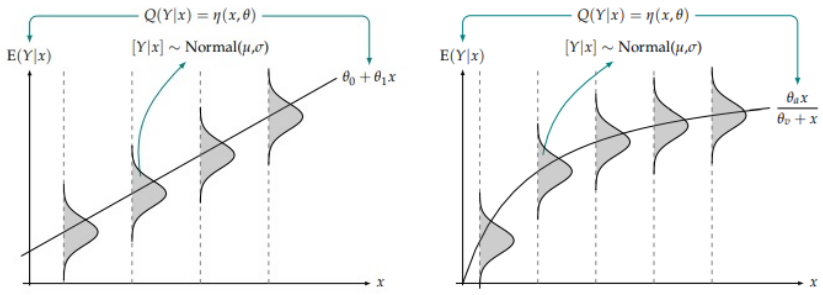
O aprendizado supervisionado tem uma referência do objeto a ser alcançado, ou seja, as classes em que o objeto deve ser classificado são conhecidas no conjunto de dados. De acordo com Pellucci *et al.* (2011), "o objetivo é que a representação seja capaz de produzir saídas corretas para novas entradas não apresentadas antes". Durante o treinamento, o algoritmo otimiza uma função de mapeamento, que é usada para classificar novas observações. Seus resultados são comparados e dessa forma é possível avaliar o desempenho do modelo. No aprendizado supervisionado podem ser criados modelos preditivos que podem ser implementados a partir de duas tarefas, a classificação (predição categórica) e a regressão (classificação numérica) (SILVA; PERES; BOSCARIOL, 2016).

3.3.1 Tarefa de regressão

A tarefa de regressão busca a relação entre uma classe e as variáveis independentes. Sendo esta segundo Souza (2020) "uma técnica estatística que estuda a relação entre uma variável resposta ou classe e uma ou mais variáveis independentes ou atributos com o objetivo de obter o melhor modelo possível que seja capaz de predizer a variável resposta". Para Zeviani, Júnior e Bonat (2013) os modelos de regressão permitem explicar o comportamento da variável dependente, bem como quantificar o quanto as variáveis independentes influenciam a variável dependente, além de selecionar os coeficientes relevantes.

Nesta tarefa, vários parâmetros são calibrados para encontrar um ajuste ótimo das variáveis independentes à variável dependente, a partir de uma curva S entre os valores de 0 e 1. A forma como a curva S é formada é determinada pela estimativa dos valores dos parâmetros de ajuste (KEMPER; VORHOFF; WIGGER, 2020). A regressão é dividida em duas categorias: regressão linear e regressão não linear. A imagem 5 ilustra os dois tipos de regressão.

Figura 5 – Regressão linear e não linear



Fonte – Zeviane (2013)

A imagem à esquerda demostra uma representação de uma regressão linear, onde a variável Y é explicada pela variável x e também pode ser explicada por desvios da relação entre essas variáveis (x,y) por uma variável z considerada aleatória. Este comportamento pode ser observado em dados reais, onde a relação linear não é perfeita. A figura à direita ilustra um modelo não-linear de Y como função de x (ZEVIANI; JúNIOR; BONAT, 2013).

De acordo com Silva, Peres e Boscariol (2016), a regressão linear estabelece que o valor da variável dependente pode ser estimado por uma combinação linear das variáveis independentes, podendo ser representada pela equação:

$$y = a + bx$$

Onde a e b são coeficientes de regressão e especificam a interceptação do eixo y a inclinação da reta. Este modelo é linear aos parâmetros a e b. Uma vez que ao menos um dos parâmetros aparece de forma não linear, o modelo é não linear e pode ser expresso pela equação:

$$y = a + bx + cx^2$$

A regressão não linear também depende de variáveis dependentes e independentes, por exemplo, e utiliza uma combinação não linear de parâmetros. Nesta equação os coeficientes a, b e c são lineares e a variável x é não linear. Em alguns casos, o modelo de uma regressão não linear pode ser transformado em um modelo linear, e assim o cálculo dos coeficientes são facilitados, este é o caso dessa equação.

3.3.2 Tarefa de Classificação

Na tarefa de classificação, o algoritmo recebe uma coleção de dados e cada instância desses dados é mapeada a um único rótulo de saída. Após a etapa de aprendizado, ocorre a etapa de teste, onde o modelo irá classificar novos dados que não estavam no conjunto de aprendizado. Por exemplo, dado um conjunto de dados de imagens, com base nos dados de treinamento é possível prever se a imagem é de um gato ou de um cão, e assim classificar em rótulos de classe como "gato"ou "cão". Dessa forma, é possível avaliar a precisão do modelo em classificar novos dados, já que a avaliação do modelo se baseia na quantidade de acertos e de erros das instâncias no conjunto de teste (MITCHELL, 1997).

A classificação pode ser dividida em dois tipos: classificação binária, onde as instancias podem ser classificadas entre duas categorias (C = 2), e a classificação multiclasse, onde as instancias podem ser classificadas num universo finito com mais de duas categorias (C > 2) (SILVA; PERES; BOSCARIOL, 2016). Para este trabalho serão utilizados os algoritmos de classificação, uma vez que se busca predizer a qual classe o aluno pertence, sendo que essas classes estão presentes na base de dados. Além disso como na base de dados a variável resposta tem mais de duas categorias, será utilizada a classificação multiclasse. Dessa forma, é necessário entender esse conceito para problemas com mais de duas categorias.

3.4 Classificação multiclasse

Existem muitos cenários em que o número de categorias que uma instância pode ser classificada é maior que dois, considerando um número finito de possibilidades. Por exemplo, em um cenário de um filme, os espectadores poderiam classifica-lo em bom, muito bom, ruim ou péssimo. Ainda considerando este cenário, poderiam também inferir qual o gênero do filme, se drama, romance, comédia. Este tipo de abordagem pertence a uma categoria de problemas na tarefa de classificação denominada classificação multiclasse, em que cada instância do conjunto de dados pertence a uma das N classes diferentes, com N maior que dois ($N > 2$) (e.g. (SILVA; PERES; BOSCARIOL, 2016) ,(SANTOS, 2017)).

Santos (2017) aponta que uma das característica desse tipo de problema é que as fronteiras de decisões são mais complexas que em problemas binários. Dessa forma alguns

classificadores como SVM precisam de ajustes para este tipo de classificação. Diante desta característica, uma forma de abordar um problema de classificação multiclasse, quando a técnica não é capaz de lidar com problemas dessa natureza, é utilizar uma abordagem de divisão deste problema em um conjunto de classificações binárias, que poderão ser resolvidas de forma singular. Quando esta abordagem é utilizada, podem ser aplicados dois métodos: Um-contra-um (One-Against-One) e Um-contra-todos (One-Against-Rest) (GALAR *et al.*, 2011).

- Um-contra-um: esta abordagem consiste em construir um classificador para distinguir um par de classes i e j, onde a classe i é o exemplo positivo e a classe j é o exemplo negativo. Durante o aprendizado, os classificadores usam um subconjunto dos dados de treinamento que contém um dos dois rótulos, enquanto as outras instancias que pertencem às outras classes, são ignoradas. Na validação, existe um padrão que será apresentado aos classificadores binários e a saída de um classificador é dada por $r_ij \in [0, 1]$ (GALAR *et al.*, 2011).
- Um-contra-todos: esta abordagem consiste em construir N classificadores binários diferentes, um para cada classe diferente. Assim, para uma determinada classe i, todos os exemplos dessa classe são considerados positivos e todos os exemplos das classes diferentes como negativos. Durante o aprendizado dos classificadores,são utilizados todos os dados de treinamento, considerando os padrões da classe única como positivos e todos os outros exemplos como negativos. Na validação, um padrão é apresentado aos classificadores binários e, em seguida, o classificador dá um resultado positivo que indica a classe de saída. Em muitos casos, a classe positiva não é única e algumas técnicas de desempates são necessárias. A abordagem mais comum usa a confiança dos classificadores para decidir a saída final, prevendo a classe a partir do classificador com a maior confiança, onde $r_i \in [0, 1]$ é a confiança para a classe i (e.g (GALAR *et al.*, 2011), (SANTOS, 2017)).

Alguns algoritmos que são utilizados para classificação binária também podem ser utilizados para a classificação multiclasse como: algoritmos de Árvores de Decisão, Naive Byes e Redes Neurais, outros como Regressão Logística e SVM precisam de ajustes para realizar esse tipo de classificação. Essas técnicas, exceto SVM, serão aplicadas a este problema de pesquisa, que pertence à categoria de problemas de classificação multiclasse. Como podem ser facilmente extendidas para esta categoria de problemas, as abordagens

de Um-contra-um e Um-contra-todos não precisarão sem empregadas. Dessa forma, não iremos aprofundar nesses tópicos.

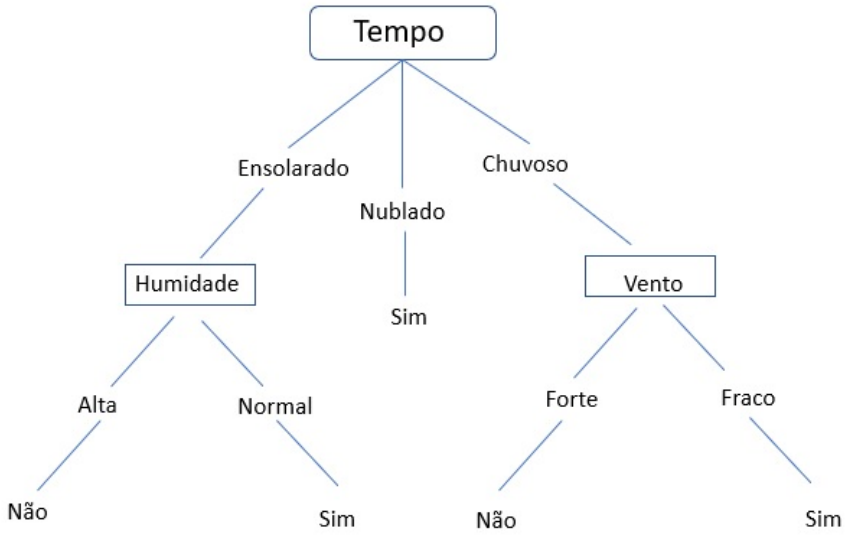
3.5 Árvores de decisão

Árvores de decisão é uma das técnicas mais utilizadas em aprendizado de máquina. Além disso, pode ser usada tanto para regressão (para respostas contínuas) quanto para classificação (para respostas categóricas). A popularidade das árvores de decisão consiste dos resultados serem facilmente interpretados, bem como o sucesso em diversas aplicações, a exemplos têm-se: diagnósticos de casos médicos; risco de crédito para análise de empréstimo. entre outras aplicações, que podem atribuir uma classificação booleana (sim ou não) ou uma classificação multiclasse, pois esta técnica pode ser facilmente utilizada para o aprendizado com mais de duas classes (MITCHELL, 1997).

As árvores de decisão classificam as instâncias da raiz até um nó folha. Este nó folha fornece a classificação da instância. Neste processo, um atributo é selecionado para ser o nó raiz da árvore, e os outros atributos servem como nós folha. Cada nó interno estabelece um teste (SE SENÃO). Com isso os atributos são divididos em subgrupos com as instâncias que alcançaram determinado ramo da árvore (e.g (SIVAKUMAR; VENKATARAMAN; SELVARAJ, 2016), (MITCHELL, 1997)).

Um exemplo de classificação através da árvore de decisão é ilustrado na Imagem 6, onde para cada nó folha é retornada a classificação associada a esta folha (neste caso, sim ou não). Esta árvore classifica se o dia está ou não favorável para jogar tênis, um exemplo clássico para demostração de funcionamento de uma árvore de decisão apresentado por Mitchell (1997).

Figura 6 – Representação de uma árvore de decisão



Fonte – adaptada de Mitchell (1997)

Para determinar qual atributo decidirá melhor os dados de destino, ou seja, qual atributo será mais útil para classificar os exemplos, o algoritmo pode usar a medida de ganho de informação. Esta medida é utilizada pelo algoritmo ID3 criado por Quinlan em 1986, onde esta medida é utilizada para selecionar os atributos candidatos em cada etapa enquanto cresce a árvore.

Mitchell (1997) afirma que para definir a precisão do ganho de informação, é preciso definir uma medida usada na teoria da informação, chamada entropia, que caracteriza a impureza de uma coleção de exemplos. Onde dado um conjunto de dados S, e contendo exemplos negativos de algum conceito alvo, a entropia de S em relação a esta classificação é calculada de acordo com a equação:

$$Entropy(S) = -P_{+}log_2P_{+} - P_{-}log_2p_{-}$$

onde P_{+} é a proporção de exemplos positivos em S e P_{-} é a proporção de exemplos negativos em S. Em todos os cálculos envolvendo entropia defini-se 0 log 0 para que seja 0. Este caso é para quando a classificação é booleana, se a classe alvo pode assumir valores maior que dois, a entropia S é definida como:

$$Entropy(S) = \sum_{i=1}^{c} -p_ilog_2p_i$$

onde p_i é a proporção de S pertencente à classe i. O logaritmo é de base 2 porque a entropia é uma medida do comprimento de codificação esperado medido em bits. Se o atributo de destino pode assumir x valores possíveis, a entropia pode ser tão grande quanto log x (MITCHELL, 1997).

3.6 Naive Bayes

O classificador Naive Bayes, assim como as árvores de decisão, também está entre as técnicas mais aplicadas no aprendizado supervisionado. O classificador é considerado um classificador ingênuo por se basear no princípio de que os atributos são independentes, ou seja a probabilidade de um evento X dado um evento Y não depende apenas da relação entre X e Y, mas também da probabilidade de observar X independentemente de observar Y (SANTOS, 2017).

O classificador naive bayes é um classificador probabilístico, que faz uso de técnicas estatísticas e cálculo de probabilidades para realizar uma classificação. O classificador utiliza o teorema de Bayes para calcular a classe mais provável de uma nova instância, calculando a probabilidade de todas as possíveis classes de um determinado conjunto de dados. Dessa forma, o algoritmo escolhe a classe com a maior probabilidade para rotular uma nova instância. O teorema de bayes é expresso por esta equação:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Onde h é a classe que se deseja classificar e D é cada nova instância do conjunto de dados. Para calcular a probabilidade de cada nova instância, se maximiza o valor de P(h|D)×P(D) e minimiza-se o valor do denominador P(h), para ocorrer a maximização de P(h|D). Em um conjunto de hipóteses H, com o intuito de determinar a hipótese mais provável de h no conjunto de dados de treinamento D, busca-se a hipótese máxima a posteriori (MAP), calculada a partir do teorema de bayes para cada probabilidade posteriori de cada hipótese, dada por:

$$argmax P(h|D) = argmax \prod_i P(a_i|h).P(h)$$

O passo a passo a ser seguido por um classificador que utiliza o teorema de bayes é: i) Se calcula a probabilidade de cada classe ocorrer; ii) se calcula a probabilidade de cada

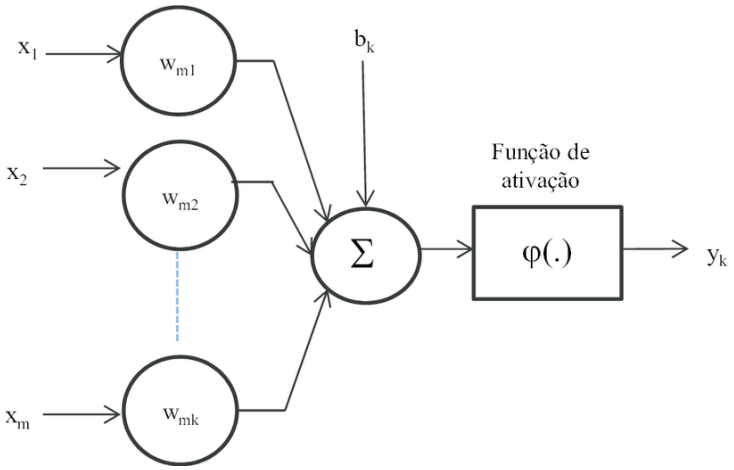
um dos atributos da instância para cada classe; iii) assim que todas as probabilidades estiverem calculadas, calcula-se a probabilidade de cada classe ocorrer utilizando a hipótese máxima a posteriori (MAP)(PARDO; NUNES, 2002).

3.7 Redes neurais artificiais

De acordo com Mitchell (1997), as Redes Neurais artificiais são sistemas distribuídos que possuem unidades de processamento chamadas de neurônios artificiais, interligadas por uma ou mais camadas. Os modelos de redes neurais artificiais são inspirados no cérebro biológico, responsável pelo processamento de diversas informações e geração de respostas. Sua criação teve como objetivo simular a capacidade de aprendizado do cérebro humano para classificar ou agrupar uma instância de um conjunto de dados. (SOUZA, 2020).

As redes neurais aprendem relacionando pesos sinápticos às conexões entre os neurônios. Elas são adaptáveis, e isso permite que sejam retreinadas ajustando esses pesos para que possa ser minimizado o erro que resulta da saída da rede. As redes neurais artificiais mais simples têm um único neurônio chamado perceptron, como pode-se observar na Imagem 7.

Figura 7 – Modelo de rede neural



Fonte – Modelo de HAYKIN apresentado por PERERA, L. C. J,2011

O perceptron recebe um vetor com os valores dos atributos x= [x1,x2..xn], que se associam ao neurônio pelo conjunto de sinapses com os pesos $w = [w_m1, w_m2...w_mk]$. A função somatória combina as entradas do neurônio e ao resultado é aplicada a função de ativação, gerando um valor de saída que representa uma de duas classes. Uma função de ativação muito utilizada é a função sigmoide, por ser capaz de lidar com as não linearidades

dos dados e lidar com alta complexidade. O perceptron também pode ter uma entrada extra, denominada "*bias*", sendo o valor sempre 1 e o peso definido por b_k. Caso ele tenha esse bias, o classificador faz o ajuste do peso b_k com o ajuste do conjunto de pesos $w = [w_m1, w_m2...w_mk]$ (SILVA; PERES; BOSCARIOL, 2016).

Durante o processo de passagem de informações da camada de entrada para a camada oculta, e em seguida para a camada de saída, ocorre a propagação direta, conhecida como *forward propagation*. dessa forma a função de ativação do neurônio é propagada para a próxima camada. Por fim, quando a informação alcança a camada de saída, é calculada a taxa de acerto do modelo, necessária para que se possa corrigir o valor dos pesos w_m utilizados no algoritmo. O cálculo para a propagação direta pode ser expresso pela equação:

$$y(x) = \sum_{i=1}^{n} x_i * w_i + bias$$

onde x_i representa cada variável de entrada no neurônio e w_i representa os pesos da rede neural.

Para utilizar uma rede neural em um problema multiclasse, uma das formas é utilizar uma rede neural do tipo multilayer perceptron, pois pode ser estendida a esse tipo de problema utilizando N neurônios binários, o que não é possível com o perceptron. Outra forma seria referenciar um vetor binário de 0 a $2^n - 1$, onde n é o número de neurônios de saída. Nesse a resposta da rede para uma instância seria um vetor que mais se aproximasse ao vetor de sua classe, utilizando o cálculo da distância de Hamming (SANTOS, 2017).

3.8 Regressão logística

A regressão logística é uma técnica estatística que tem sido muito utilizada em aplicações na área da saúde, administração, educação entre outros. O modelo consegue expressar uma relação entre a variável dependente e uma ou mais variáveis independentes. Dessa forma a regressão logística considera uma relação entre as variáveis independentes e a razão de chance da variável dependente. Conforme Assis (2018) a regressão logística também consegue estimar a probabilidade de fracasso e de sucesso dos atributos, podendo ser representada pela equação:

$$E(Y|x) = G(X) = G(\beta_0 + \beta_i X_i)$$

onde G(X) é uma função cujos valores variam entre 0 e 1, Y é a variável dependente, β é o vetor dos coeficientes desconhecidos das variáveis independentes, e X é o vetor das instâncias.

Para a regressão logística ser aplicada a um modelo de classificação multiclasse, a técnica precisa de alguns ajustes fazendo uso da regressão logística multinomial, que é modelada a partir do ajuste simultâneo da quantidade de categorias k-1 modelos. Dessa forma, são estimados k-1 vetores dos parâmetros βi, correspondentes a k-1 categorias da variável dependente. Assim têm-se k-1 comparações com a categoria escolhida.

3.9 Métricas para avaliação de um modelo de aprendizado de máquina

A avaliação do desempenho de um modelo de aprendizado de máquina é uma importante etapa no processo de escolha do melhor classificador para a predição de novas instâncias. Dessa forma, o cálculo da avaliação é realizado quando o modelo classifica instâncias até então desconhecidas no conjunto de dados, geralmente essas instâncias estão em um conjunto de teste, que é dado como entrada para o algoritmo após o processo de aprendizado. A partir disso é possível calcular diferentes métricas de avaliação, utilizando a matriz de confusão.

A matriz de confusão é uma tabela com m x m dimensões, em que m é o número de classes a serem preditas, representando as colunas e as linhas da tabela. Cada célula da matriz contém o número de instâncias classificadas correta ou incorretamente para cada classe. As classificações corretas contabilizam os valores que estão na diagonal da matriz, como apresentado na Figura 8 (SILVA; PERES; BOSCARIOL, 2016).

Podemos visualizar na Figura 8 quatro medidas que servem como base para o cálculo das métricas de avaliação de um modelo:

- verdadeiro Positivo (VP): quantidade de instâncias positivas classificadas como positivas.
- verdadeiro Negativo (VN): quantidade de instâncias negativas classificadas como negativas.

Figura 8 – Matriz de confusão

	Valor predito	
Valor Real	Sim	Não
Sim	Verdadeiro positivo (VP)	Falso Negativo (FN)
Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte – Jailma Januário, 2021

- Falso Positivo (FP): quantidade de instâncias negativas classificadas incorretamente como positivas.
- Falso Negativo (FN): quantidade de instâncias positivas classificadas incorretamente como negativas.

A partir desses valores é possível calcular outras métricas importantes na avaliação dos modelos:

- Acurácia: uma métrica muito utilizada em aprendizado de máquina para saber o percentual de acerto do modelo, É calculada pela razão do total de instâncias corretamente classificadas (positivo e negativo) pelo número total de instâncias, conforme a equação:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

- Sensibilidade: esta métrica indica o percentual das observações positivas classificadas corretamente. O cálculo dessa métrica é dado pela razão seguinte equação.

$$Sensibilidade = \frac{VP}{VP + FN}$$

- Especificidade: esta métrica indica o percentual de observações negativas que foram classificadas corretamente. O cálculo da especificidade é dado pela equação

$$Especificidade = \frac{VN}{VN + FP}$$

- F1 Score: é uma métrica de desempenho para modelos de classificação e é calculada como a média harmônica de precisão e Sensibilidade.

$$F1Score = \frac{2 * precisão * Sensibilidade}{precisão + Sensibilidade}$$

Embora essas métricas sejam muito utilizadas em problemas de classificação binária, também podem se utilizadas para a classificação multiclasse, abordando algumas estratégia como macro-média e micro-média (BARROS, 2018).

- Macro-média: para o cálculo da macro-média deve-se realizar a soma dos resultados por cada classe dividido pelo número da classe, pode ser representado pela equação:

$$MaM_{precisão} = \sum_{i=1}^{n} \frac{precisão_n}{n}$$

- Micro-média: para calcular a micro-média deve-se realizar as contagens dos resultados das classificações (falsos positivos, falsos negativos) que compõem o cálculo da métrica para cada classe analisada. Por exemplo, para realizar o cálculo da micro-média da sensibilidade, tendo as classes x e y

$$MiM_{sensibilidade} = \frac{VP_x + VP_y}{VP_x + VP_y + FN_x + FN_y}$$

Para este trabalho serão utilizadas as métricas de acurácia e macro médias das medidas de sensibilidade e especificidade, uma vez que avaliar apenas a acurácia para um problema de classificação não é ideal, pois essa métrica calcula todos os acertos tanto negativos como positivos, o que pode levar o modelo a ter 90% de acurácia, isso não acontece no mundo real principalmente quando o problema apresenta uma característica que é o desbalanceamento de classes. O desbalanceamento de classes ocorre quando uma determinada classe tem uma representatividade no conjunto de dados do que outra.

3.10 Teste estatístico

Os dados que resultaram dos experimentos podem ser submetidos a vários tipos de análise estatística como a análise de variância (ANOVA) para verificar se há uma diferença significativa entre os classificadores e entre as classes do conjunto de dados, ao se constatar essa diferença pode-se aplicar um teste estatístico de comparação entre as médias, essa comparação pode ser realizada aplicando a comparação entre pares, dessa

forma é necessário elaborar hipóteses para serem comparadas. As hipóteses da ANOVA são:

- H_0 Não existem diferenças significativas entre as médias(hipótese nula);
- H_1 Existem diferenças significativas entre as médias (hipótese alternativa);

Para a comparação entre pares será utilizada o Teste de Tukey, este teste foi proposto por Tukey para problemas com erro do tipo I. Um erro do tipo I ocorre quando se rejeita a hipótese nula (H_0) e conclui que existem diferenças significativas entre as médias, quando de fato, não existem. Segundo Mendes (2017), o teste de Tukey é bastante popular por ter um rigor significativo nos resultados e aplicação fácil, pode ser demostrado pela equação:

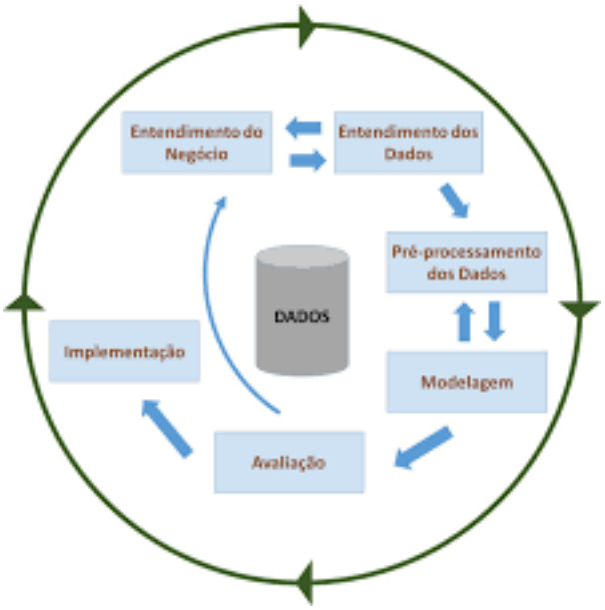
$$\Delta\alpha = q(\alpha)\sqrt{\frac{QMR}{r}}$$

onde $q(\alpha)$ é o valor tabelado para Tukey, com base no número de tratamento e nos graus de liberdade do resíduo, α é o nível de significância, QMR é a estimativa do desvio padrão residual e r é o número de repetições por tratamentos. Para determinar se há diferenças, verifica-se o valor da diferença entre as duas médias, se for igual ou maior que a diferença mínima significativa (Δ) Mendes (2017).

4 Metodologia

Nesta Seção, é apresentada a metodologia utilizada para o desenvolvimento desta pesquisa. Conforme mencionada na Seção 1, a estrategia metodológica utilizada é a CRISP-DM (CRoss-Industry Standard Process for Data Mining), uma metodologia sólida desenvolvida com base nas experiências práticas e reais de como as pessoas conduzem projetos de mineração de dados nas industrias (CHAPMAN *et al.*, 2000). Essa metodologia consiste em seis etapas: entendimento do negócio; entendimento dos dados; pré-processamento dos dados; modelagem; avaliação e implementação, conforme ilustrado na figura 9. Nela, as setas indicam que em qualquer etapa da metodologia em que o projeto estiver, sempre é possível voltar à etapa anterior.

Figura 9 – Modelo CRISP-DM



Fonte – Coelho et all., 2016

4.1 Entendimento do negócio

Esta fase consiste em entender a problemática para que possam ser definidos os objetivos que regem o desenvolvimento do projeto, Nas Seções 1 e 2 está descrita detalhadamente esta problemática. A seção 1 introduz o problema da evasão no ensino superior e como isso pode afetar a sociedade. Assim, foi possível definir quais eram os objetivos a serem alcançados na realização desta pesquisa. A Seção 2 descreve uma revisão

da literatura com princípios sistemáticos, com foco em trabalhos que tratam da evasão do ensino superior e algumas técnicas que são utilizadas para auxiliar as instituições a desenvolverem ações mais assertivas para a solução dessa problemática.

4.2 Entendimento dos dados

Esta etapa permite que os pesquisadores se familiarizem com os dados que serão utilizados e consiste em: coleta inicial dos dados; descrição dos dados; exploração e verificação da qualidade dos dados.

Os dados foram coletados no site do INEP, uma autarquia federal vinculada ao MEC (ministério da educação), o qual realiza anualmente o censo da educação superior e também divulga anualmente informações sobre todas as instituições brasileiras e seus cursos de graduação presencial e a distância, também possui informações de todos os estudantes que ingressaram no ensino superior desde 1995. No site é possível obter relatórios sobre o ensino superior e os microdados, sendo este o menor nível de desagregação de dados recolhidos por suas pesquisas para que pesquisadores, professores, e toda a população possam realizar suas próprias análises sobre os dados. Dessa forma, torna-se possível realizar a pesquisa sobre evasão no ensino superior no Brasil.

O ano escolhido para a realização da análise para esta pesquisa foi o ano de 2018, uma vez que, até então eram os dados mais recentes que estavam disponíveis no site. Esta base de dados contém mais de 12 milhões de linhas com 99 variáveis quantitativas e qualitativas, com informações de estudantes na modalidade presencial e a distância. Junto ao arquivo que contém informações dos alunos, o INEP também disponibiliza um dicionário de variáveis, onde é explicada cada informação presente na base de dados, e também a partir de que ano aquela informação começou a estar presente no conjunto de dados.

4.2.1 Situação do aluno no curso

Cada aluno possui apenas um tipo de vínculo com o curso no ano em que o Censo é realizado. Estes vínculos são:

- Formado: situação do aluno que, no ano de realização do censo, concluiu o seu curso;

- Cursando: situação do aluno que, no ano de realização do censo, está com a matrícula ativa no curso;
- Matrícula trancada; situação do aluno que, no ano de realização do censo está com a matrícula desativada no curso.
- Desvinculado do curso: situação do aluno que, no ano de realização do censo está com a matrícula inativa no curso.
- Transferido para outro curso da mesma instituição: situação do aluno que, no ano de realização do censo está com a matrícula ativa em outro curso da mesma instituição;
- Falecido: situação do aluno que estava falecido no ano de realização do senso.

Dessas categorias foram selecionadas quatro para serem preditas no modelo de aprendizado de máquina, utilizando classificação multiclasse, pois cada aluno pode estar apenas em uma dessas categorias. Sendo assim, o modelo construído neste trabalho irá classificar o aluno entre Formado, Matrícula trancada, Desvinculado do curso ou Transferido para outro curso da mesma instituição.

A base de dados contém informações referentes ao ano de realização do censo, sexo, cor/raça, idade, local de nascimento, local onde estuda, tipo de deficiência ou superdotação do aluno. Ela traz também informações sobre o aluno ser assistido por algum programa de apoio social como: auxílio moradia, auxílio transporte, creche, alimentação, se realizava algum estágio, projeto de extensão, pesquisa ou ensino. A coleta realizada pelo Censo da Educação Superior (CES) é realizada em etapas e garante a qualidade dos dados apresentados pois na última etapa é verificada a qualidade da informação fornecida pelas Instituições de Ensino Superior(ASSIS, 2018).

4.3 Pré-processamento dos dados

Nesta etapa é realizado o pré processamento dos dados, que consiste na seleção, limpeza e transformação dos dados que vão ser utilizados. Durante esse processo é possível detectar e corrigir dados incorretos, tratar variáveis com valores faltantes, selecionar variáveis relevantes e, se for preciso, aplicar alguma técnica de normalização dos dados. Dessa forma, um estudo para cada variável se faz necessário, para saber qual processo precisa ser realizado, para que ao final o conjunto de dados esteja pronto para os algoritmos de classificação.

4.3.1 Seleção da população e variáveis de interesse

Como o conjunto de dados disponibilizados pelo INEP contém informações de alunos que estudam em instituições públicas e privadas, nas modalidades presencial e a distância, em um primeiro momento foi preciso selecionar a população de interesse. Foram selecionadas informações de alunos que estudam na modalidade presencial e em instituições públicas. Após essa identificação foram selecionados os dados que continham as classes de interesse para a pesquisa: formado, matrícula trancada, desvinculado do curso e transferido para outro curso da mesma instituição. Após essa seleção inicial o conjunto de dados passou a ter 3.381.160 observações.

Os atributos foram selecionados conforme a pesquisa bibliográfica, considerando fatores institucionais, acadêmicos, socioeconômicos, demográficos, pessoais/sociais, anterior ao ingresso do aluno no curso, como tipo de escola de conclusão do ensino médio. Também foram selecionados os dados de acordo com o trabalho de Teodoro e Kappel (2020), que fez aplicações de técnicas de aprendizado de máquina em busca de descobrir os melhores atributos para classificar um aluno em risco de evasão utilizando os dados disponíveis pelo INEP. Assim foram selecionadas 46 características para o estudo, e após a remoção de algumas variáveis e instâncias, o motivo será melhor explicado abaixo, o conjunto de dados passou a ter 2.525.411 linhas e 26 variáveis. Desse total foi gerada uma amostra estratificada de 500.000 observações para compor o conjunto de dados final.

4.3.2 Valores faltantes

Existem algumas formas de lidar com valores faltantes em um conjunto de dados, como a imputação de valores utilizando a média, mediana e a moda (valor mais frequente em um determinado conjunto) para variáveis numéricas, e a criação de uma nova categoria para variáveis categóricas, pode-se optar pela exclusão da variável, ou instância que apresente valores faltantes. Outra possibilidade é utilizar algoritmos de aprendizado para predizer o valor faltante, como algoritmos de regressão e kNN (DONG; PENG, 2013).

Neste estudo optou-se por excluir as variáveis que apresentassem mais de 50% dos dados faltantes, conforme a Tabela 4 as variáveis que continham esse quantitativo foram retiradas da base, uma vez que a proporção de dados ausentes está relacionada com a

qualidade de inferências estatísticas, e apesar da literatura não apresentar um quantitativo específico de dados faltantes para a retirada de uma variável, pois vai depender de cada problema. Dong e Peng (2013) apontam que a análise estatística tende a ser tendenciosa quando mais de 10% dos dados são ausentes.

Tabela 4 – Variáveis excluídas por conter muitos valores faltantes

Variáveis	***Percentual de valores faltantes***
Reserva de vagas por etnia, deficiência, ensino público, renda familiar, outra categoria	95.7%
Apoio alimentação, bolsa permanência,bolsa de trabalho, apoio a material didático, apoio moradia, apoio transporte	96.1%
Estágio complementar, extensão,monitoria, pesquisa	88.1%
Bolsa estágio,	97.1%
Bolsa extensão,	91.2%
Bolsa monitoria	98.1%
Bolsa pesquisa	97.8%
Aluno parfor	87.7%

Fonte – Jailma Januário, 2021

As variáveis que tinham menos de 50% de ausência de informações foram retiradas, como o caso da variável de identificação do estado de nascimento que apresentou 25.3% dos dados faltantes.

4.3.3 Conjunto de dados final e análise descritiva dos dados

O conjunto de dados final é composto por 500.000 linhas e 26 atributos. Este quantitativo foi gerado a partir de uma amostra estratificada, mantendo as proporções das categorias a serem preditas do conjunto de dados. Essa técnica de amostragem foi escolhida por garantir que cada categoria tivesse representatividade no conjunto de dados, já que em uma amostragem simples algumas das categorias poderiam não ser incluídas na amostra. A amostragem estratificada divide o número de população em grupos (estratos) que mantém a representatividade da população, ou seja, a proporção de cada estrato

na amostra deve ser a proporção de cada grupo na população (BOLFARINE; BUSSAB, 2004). Os estratos foram gerados utilizando as categorias para que as mesmas tivessem a proporção igual nos dois conjuntos de dados.

A Tabela 5 a seguir apresenta os atributos selecionados para compor o conjunto de dados final, de acordo com os fatores que levam o aluno a desistir do curso encontrados na revisão sistemática da literatura e no trabalho de Teodoro e Kappel (2020).

Tabela 5 – Fatores e suas variáveis analizadas

Fatores	***Variáveis pertencentes***
Acadêmicos	Turno , código do curso, carga horária total, carga horária integrada, semestre de referência, atividade extracurricular, data de ingresso no curso
Socioeconômicos	Apoio social
Demográficos	Estado de nascimento, Nacionalidade
Pessoais	Cor-raça, sexo, idade
Anterior ao ingresso do aluno no curso	Ingresso vestibular, enem, avaliação seriada, seleção simplificada, vaga remanescente,programa especial, transferência exofício, ingresso- egresso, reserva de vagas: Étnica; Deficiência; Ensino público; Renda familiar; Outro tipo de reserva de vagas. Tipo de escola que concluiu o ensino médio. Ingresso por convênio, decisão judicial

Fonte – Jailma Januário, 2021

4.3.4 Análise descritiva dos dados

A análise descritiva dos dados faz uso de métodos estatísticos para descrever e resumir um conjunto de dados. Também com ela podemos identificar anomalias presentes nos dados e dessa forma tratar estas anomalias. Além disso, também é possível através da análise descritiva entender como as variáveis se relacionam entre si, e como estão distribuídas no conjunto de dados (REIS; REIS, 2002).

Com relação as variáveis categóricas são apresentadas as frequências no conjunto de dados, para as variáveis numéricas são apresentadas as medidas de tendência central: moda, mediana e a média. Nesta subseção serão apresentadas apenas algumas variáveis, que em outros trabalhos (e.g (PEREZ; CASTELLANOS; CORREAL, 2018),(KEMPER; VORHOFF; WIGGER, 2020),(PEREIRA; ZAMBRANO, 2017) (TEODORO; KAPPEL,

2020)) são demostradas como atributos comuns no perfil do aluno que evade, com relação às categorias: desvinculado; transferido; formado e matrícula trancada.

Após a preparação dos dados, foi possível realizar uma análise descritiva das variáveis independentes e como elas se relacionam com a variável a ser predita. O conjunto de dados é composto por variáveis numéricas e categóricas. A Tabela 6 apresenta, para cada fator, o quantitativo de variáveis por tipo. Pode-se perceber que o conjunto de dados possui majoritariamente variáveis do tipo categórica 22 (binárias e multinomial) ao total e 4 variáveis numéricas discreta.

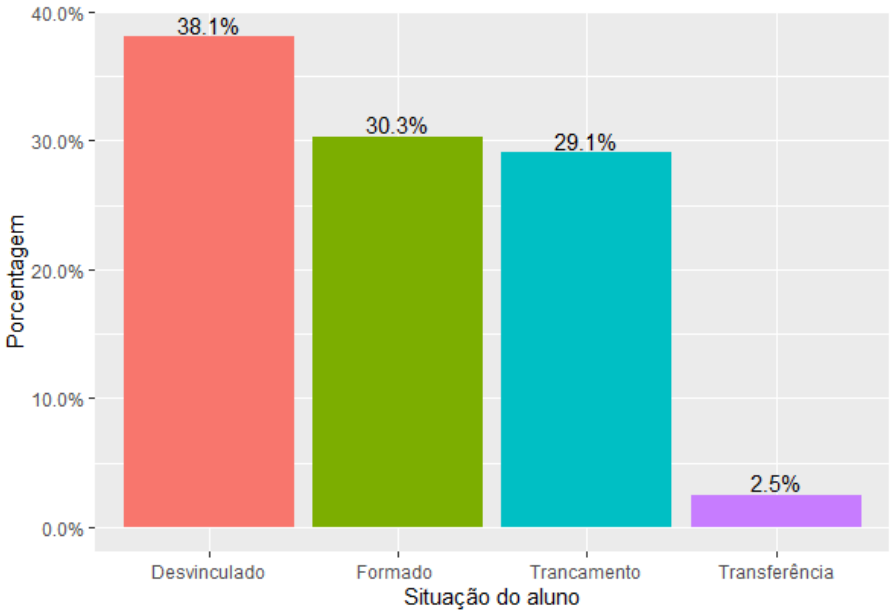
Tabela 6 – Quantidade de variáveis por fatores e tipo

Fatores	***tipo de variável***	**quantidade**
Acadêmicos	Numéricas	2
Acadêmicos	Categóricas	5
Socioeconômicos	Numéricas	0
Socioeconômicos	Categóricas	1
Demográficos	Numéricas	0
Demográficos	Categóricas	2
Pessoais	Numéricas	1
Pessoais	Categóricas	2
Anterior ao ingresso no curso	Numéricas	1
Anterior ao ingresso no curso	Categóricas	12

Fonte – Jailma Januário, 2021

A Figura 10 apresenta o percentual para cada categoria matrícula trancada, desvinculado, transferido, formado.

Figura 10 – Percentual de situação do aluno



Fonte – Jailma Januário, 2021

Observa-se que o maior percentual de alunos na base de dados está entre os desvinculados (38.1%) e os formados (30.3%). Além disso, percebe-se que na dimensão de evasão temporária apresentada por Silva, Cabral e Pacheco (2016) tem-se o percentual de 29.1%. Nesse caso, em que o aluno interrompe o seu ciclo de estudo (Trancamento), ele pode ou não voltar para o curso, após um determinado período. Por outro lado, o percentual de alunos que está na categoria de transferência de curso para a mesma instituição é de 2.5%. Isso demostra uma desigualdade na distribuição das categorias. Neste estudo optou-se por não ser aplicada nenhuma técnica de balanceamento das classes. Dessa forma os algoritmos foram treinados com o quantitativo real das classes no conjunto de dados. A Tabela 7 apresenta a frequência absoluta e relativa das categorias presentes na base de dados.

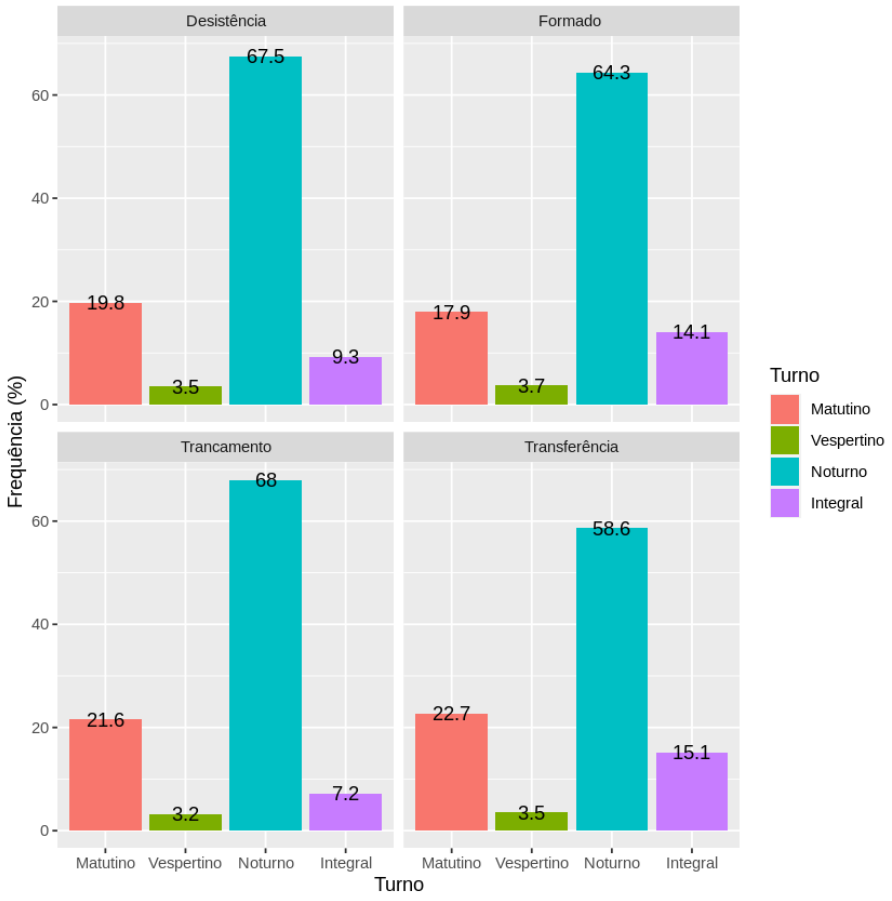
Tabela 7 – Distribuição de frequências segundo a situação

Situação	***Frequência absoluta***	***Frequência relativa (%)***
Matrícula trancada	145500	29.1
Desvinculado do curso	190500	38.1
Transferido	12700	2.5
Formado	151300	30.3

Fonte – Jailma Januário, 2021

A Figura 11 representa a relação da situação de aluno por turno. No sistema de ensino, os cursos estão disponíveis em três turnos: matutino, vespertino, noturno e integral. Como pode-se observar os maiores índices de Desistência, Trancamento e transferência estão no período noturno com 67%, 68.0%, 58,6%, respectivamente, bem como a porcentagem de formados (64.3%).

Figura 11 – Porcentagem de situação do aluno por turno



Fonte – Jailma Januário, 2021

A fim de entender por que o percentual para todas as categorias foram maiores para o turno da noite, a Tabela 8 a seguir apresenta o quantitativo para cada turno. Pode-se concluir que a representação na base de dados dos alunos que estudam no turno noturno é maior que as outras categorias de turno.

Tabela 8 – Quantitativo de alunos por turno e categorias

Turno/categoria	Desvinculado		Transferência		Trancamento		Formado	
	Frequência							
	Absoluta	Relativa(%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Matutino	37677	19.8	2884	22.7	31413	21.6	27113	17.9
Vespertino	6623	3.5	450	3.5	4648	3.2	5570	3.7
Noturno	128570	67.5	7442	58.6	98972	68.0	97344	64.3
Integral	17630	9.3	1924	15.1	10467	7.2	21273	14.1

Fonte – Jailma Januário, 2021

Dos alunos que evadem, diversas variáveis podem causar essa desistência, uma delas é que geralmente no turno da noite estudam pessoas que trabalham, (MARANHãO; VERAS, 2017) apresentam o estudo nesse período como uma alternativa para pessoas que possuem alguma atividade laboral, ressaltando que o estudo nesse período não é exclusivamente para essas pessoas, mas que a relação trabalho e estudo é um elemento comum aos estudantes desse período. De acordo com Silva, Cabral e Pacheco (2016) a conciliação entre trabalho e estudo é um dos fatores que levam o aluno a evadir.

Os dados do INEP também contêm informações do sexo dos alunos. Dessa forma, foi realizada uma análise da variável sexo. A Figura 12 apresenta os resultados, podemos perceber que a maioria dos alunos que estão classificados como Desvinculado, Trancamento, Transferência e Formado são do sexo feminino representando 50.4%, 50.9%, 51.1.% e 59.5%, respectivamente. Analisando a Tabela 9 a seguir, com as frequências absolutas e relativas, pode-se perceber que a representatividade feminina no conjunto de dados é maior que a masculina, Isto pode explicar a predominância da evasão do sexo feminino entre as categorias.

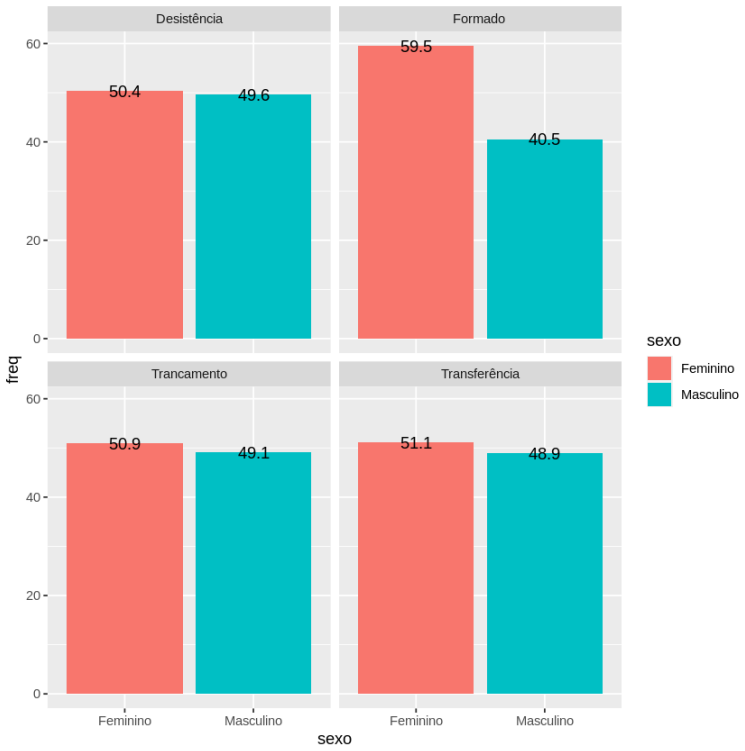
Tabela 9 – Quantitativo de alunos por sexo e categorias

Sexo/categoria	Desvinculado		Transferência		Trancamento		Formado	
	Frequência							
	Absoluta	Relativa(%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Feminino	95922	50.4	6490	61.1	74076	50.9	90004	59.5
Masculino	94578	49.6	6210	48.9	71424	49.1	61296	40.5

Fonte – Jailma Januário, 2021

Estas medidas, contudo se contrapõem ao trabalho de Andrade (2016), que apresenta um maior índice de evasão para o sexo masculino, mesmo a entrada de novos alunos estando bem distribuída entre eles.

Figura 12 – Porcentagem de situação do aluno por sexo



Fonte – Jailma Januário, 2021

Com relação à idade dos alunos, sendo a variável idade uma variável contínua, a Tabela 10 mostra as estatísticas descritivas da idade no conjunto de dados para cada situação. Em média, a idade para os alunos que estão nas categorias desistência, formado e trancamento são 27.49, 27.81 e 27.75, respectivamente. Já para a categoria Transferência, a média de idade é de 25.29. A idade mais frequente para as situações são 21 anos, para a categoria transferência e desistência, e 22 para evasão temporária (trancamento) e 23 para a categoria de formados.

Tabela 10 – Estatísticas da variável idade por categoria

	Min.	1st Qu.	Mediana	Média	3rd Qu.	Max.	Moda
Desistência	16.00	22.00	25.00	27.49	31.00	85.00	21
Formado	18.00	23.00	25.00	27.81	30.00	83.00	23
Trancamento	16.00	22.00	25.00	27.75	31.00	85.00	22
Transferência	17.00	21.00	23.00	25.29	28.00	72.00	21

Fonte – Jailma Januário, 2021

De acordo com a literatura, outro ponto a se observar com relação aos alunos que estão no ensino superior é a forma de ingresso nas universidades. Existem várias formas de ingresso no ensino superior. O aluno pode ingressar por meio de vestibular, ENEM, avaliação seriada, simplificada, entre outros tipos. A Tabela 11 a seguir apresenta o quantitativo de ingressantes por diferentes formas. Como podemos observar o maior número de ingressantes é por vestibular. O gráfico abaixo apresenta este tipo de ingresso por categoria.

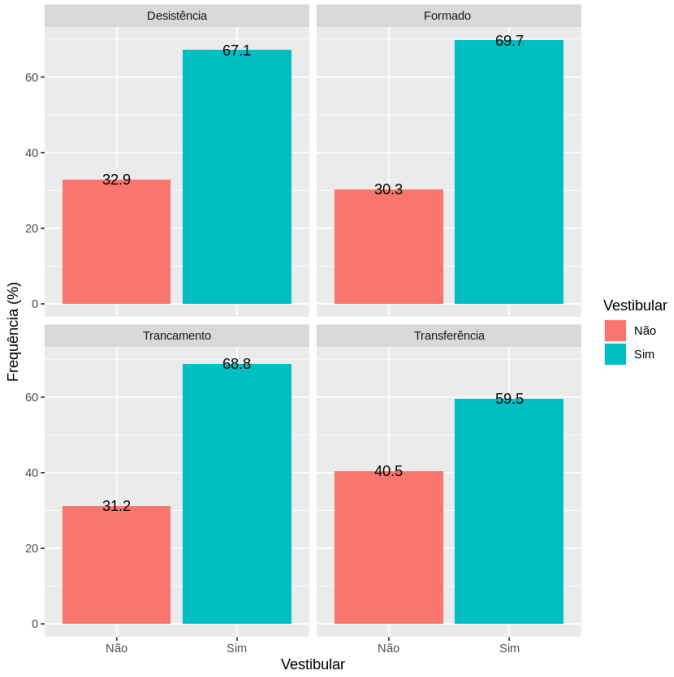
Tabela 11 – Número de ingressantes por forma de ingresso

Formas de ingresso	***Quantidade***
Vestibular	341054
Enem	84697
Reserva de vagas	19861
Exofício	873
Ingresso-egresso	344
Vaga em programa especial	1555
Vaga remanescente	64578
Avaliação seriada	2026
Seleção simplificada	20803

Fonte – Jailma Januário, 2021

A Figura 13 apresenta a porcentagem de alunos por tipo de ingresso com relação a variável vestibular. Como podemos observar, a maioria dos estudantes que estão na

Figura 13 – Tipo de ingresso por vestibular no ensino superior



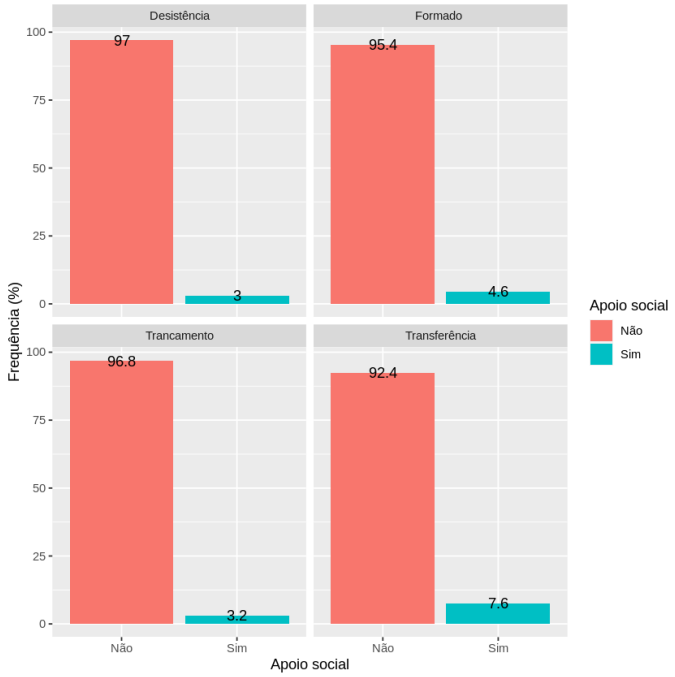
Fonte – Jailma Januário, 2021

categoria desvinculado, Trancamento, Transferência e formado (67.11%, 68.8%, 59.5%, 69.7% respectivamente) ingressaram no ensino superior por meio desse exame.

Um dos principais fatores que influenciam na permanência do aluno no curso é a questão socioeconômica. Dessa forma, é preciso entender se os alunos que estão na situação de transferido, desvinculado, formado ou transferido para outro curso da mesma instituição, possuem algum tipo de apoio social que as instituições públicas oferecem. A Figura 14 representa alunos que receberam ou não algum tipo de apoio social, os quais tendem a abandonar, ou transferir do curso. Em contrapartida esse número cai visivelmente se o aluno possui algum tipo de apoio social, reafirmando o fato de que os programas de incentivo financeiro atuam diretamente para mudar esse cenário.

Neste tópico foi apresentada uma descrição de cinco variáveis preditoras, uma variável numérica e quatro variáveis categóricas das que compõem a base de dados. Com os gráficos e tabelas apresentados foi possível analisar a distribuição (relativa e absoluta) das variáveis turno, sexo, tipo de ingresso e apoio social, e as medidas de tendência central da variável idade, com relação as categorias: desvinculado; formado; trancamento; e transferido para outro curso da mesma IES.

Figura 14 – Apoio social por situação no curso



Fonte – Jailma Januário, 2021

4.4 Modelagem

Nesta fase do projeto são selecionadas e implementadas as técnicas escolhidas. Em algumas modelagens pode ser preciso voltar a etapa de pré-processamento, para que o conjunto de dados esteja adequado ao algoritmo escolhido. As técnicas que foram implementadas são: Árvore de Decisão, Naive Bayes, Regressão Logística e Redes Neurais Artificiais, os detalhes dessas técnicas estão descritos na Seção 3.

As implementações foram realizadas utilizando o pacote caret (classification and regression training) da linguagem de programação R, uma linguagem muito utilizada para a manipulação, análise e visualização dos dados. Este pacote foi utilizado por possuir várias implementações de algoritmos de aprendizado de máquina e funções para divisão do conjunto de dados, ferramentas de pré-processamento, entre outras funcionalidades.

Para a aplicação das técnicas, em um primeiro momento foi realizado o particionamento do conjunto de dados em um conjunto de treinamento, com 75% dos dados, e um conjunto de teste, com 25% dos dados. A divisão foi realizada de forma estratificada para manter as mesmas proporções nos dados de treinamento e teste iguais às do conjunto de dados original.

Os algoritmos foram treinados com o conjunto de dados de treinamento, utilizando a validação cruzada, uma técnica que "faz reamostragens para realizar o treinamento de um modelo"(ASSIS, 2018). Na validação cruzada, os dados são divididos em um número fixo (k) de subconjuntos. O modelo é treinado em todos menos em um (k-1), e em seguida o modelo é avaliado com o conjunto de dados que não foi usado no treinamento. Esse processo é repetido k vezes. Os números de k, mais utilizados na literatura são 3, 5 e 10. Pode-se utilizar outros valores para o tamanho de k, no entanto um valor muito grande aumenta o custo computacional da técnica. De acordo com Cunha (2019), 10 é o valor ideal para obter uma melhor estimativa de erro. Dessa forma foi selecionado esse valor de k, para este trabalho.

Após o processo de treinamento o modelo recebe como entrada os dados do conjunto de teste. A partir da classificação das instâncias deste conjunto de dados, que não foram vistas durante o processo de treinamento, é possível medir a capacidade de generalização do modelo.

Para a implementação do algoritmo Naive Bayes, foi utilizado o pacote e1071 que está incluso no pacote Caret. Esta implementação assume a independência entre as instâncias do conjunto de dados e a classe de interesse. O algoritmo permite alterar poucos parâmetros, entre eles está o parâmetro *laplace*, utilizado para realizar uma regularização através de suavização de *Laplace*. Por padrão esse valor é 0, o algoritmo utilizado foi executado com as configurações padrão do pacote, assumindo dessa forma uma distribuição gaussiana.

Para a técnica de árvore de decisão foi utilizada a implementação do algoritmo C5.0 (no Caret é o pacote "C50"). Esta implementação pode ser treinada com um método *ensemblle (boosting)*, configurada através do parâmetro *trials*. No entanto esta configuração não foi feita na implementação para este trabalho. O algoritmo também permite escolher se o modelo é baseado em regras ou em árvore de classificação. Com isso foi preciso configurar o parâmetro *model*, para que o modelo treine árvores de decisão. O atributo *winnow* (por padrão é falso) especifica se deve ser realizada uma pré-seleção de variáveis antes da criação do modelo.

A técnica de Redes Neurais artificiais foi utilizado o método "nnet"do pacote caret, que permite a implementação de redes neurais artificiais *multilayer perceptron* com uma camada escondida. Por padrão, a implementação no pacote possui dois parâmetros configuráveis: decaimento (decay) dos pesos, que serve para realizar a regularização do

modelo, e a quantidade de neurônios na camada oculta (syze), que recebe os valores 1,3 e 5 por padrão. Além desses parâmetros, também podem ser configurados os parâmetros *MaxNWts* (número máximo de pesos permitido) e *maxit* (número máximo de iterações) nos quais foram utilizados os valores padrão 1.000 e 100, respectivamente.

A técnica de regressão logística foi implementada utilizando o pacote "nnet"presente no caret. Para esta implementação foi preciso configurar o parâmetro *family* para multinomial, uma vez que, a variável de interesse possui mais de duas categorias. A função *logit* também foi utilizada para garantir que o modelo produza uma probabilidade.

4.5 Avaliação

Nesta etapa é definida e realizada a avaliação de desempenho dos modelos implementados. Baseado no que foi identificado na literatura como métricas mais utilizadas para este trabalho, foram selecionadas as métricas de acurácia, sensibilidade e especificidade, detalhadas na seção 3. A principal métrica a ser utilizada para a escolha do melhor classificador será a sensibilidade, pois com ela é possível medir a capacidade de um modelo classificar corretamente os valores positivos, ou seja quem está corretamente nas classes: desvinculados, matrícula trancada e transferido do curso. A escolha foi baseada em um estudo realizado por Assis (2018) onde destaca que não poder dar assistência a um aluno que está em risco de evadir ou trancar a matrícula tem um custo maior do que dar assistência ao aluno que não está em risco.

5 Resultados

Nesta seção serão apresentados os resultados da avaliação de cada modelo treinado com base nas métricas de avaliação acurácia, sensibilidade e especificidade. Por fim, é realizada uma análise de variância para verificar se há diferenças estatísticas entre os modelos e as diferentes classes do conjunto de dados. A avaliação dos modelos preditivos tem por finalidade verificar a capacidade de generalização do classificador para novos dados. Dessa forma, as medidas de desempenho são calculadas com base no conjunto de teste, que possui os dados que não foram utilizados no treinamento.

Para cada algoritmo é apresentada a medida de acurácia. No entanto, esta métrica não é o suficiente para classificar um modelo como o melhor quando existe uma distribuição desbalanceada das classes nos dados, como o caso deste trabalho. Dessa forma, também são calculadas as macro médias de sensibilidade e especificidade. Os resultados serão descritos para cada classificador, e logo após, será feita uma comparação entre os modelos. Também nesta seção será demostrada uma análise de variância para cada classe predita e para os modelos.

5.0.1 Avaliação dos modelos

O classificador Naive Bayes apresentou um valor de acurácia de 0.62, considerado moderado, já que o valor ideal é próximo de 1.0, e 0.46 para a macro média da sensibilidade, como podem ser observado na Tabela 20, indicando a proporção geral do modelo para os verdadeiros positivos. As métricas foram calculadas utilizando a matriz de confusão do modelo na Tabela 12. Como pode-se observar a classe com o maior número de instâncias classificadas corretamente é a classe D (desvinculado), que possui uma maior representatividade no conjunto de dados.

Na tabela 13 podem ser visualizados os valores das métricas de sensibilidade e especificidade para cada classe, indicando um valor de 0.81 para a sensibilidade com relação à classe D (desvinculado), seguido pela classe F (formado) com um valor de 0.87, isso leva a concluir que o modelo retorna a maioria dos resultados positivos para estas classes. Já para a classe T (trancamento) e TC (transferência) o modelo apresentou um valor baixo para a sensibilidade, 0.16 e 0.0 respectivamente para esta métrica.

Tabela 12 – Matriz de confusão Naive bayes

Classe prevista				
Classe real	T	D	TC	F
T	5907	3919	498	455
D	26510	38661	2257	4124
TC	0	0	0	2
F	3958	5045	418	33246

Fonte – Jailma Januário, 2021

Com relação à especificidade, o modelo apresenta uma proporção moderada (0.57) da proporção dos verdadeiros negativos para a classe D (desvinculado), para as demais o modelo apresenta valores altos a partir de 0.89, indicando que para as outras classes o classificador retorna a maioria dos verdadeiros negativos.

Tabela 13 – Medidas de desempenho por classe modelo Naive Bayes

	T	**D**	**TC**	**F**
Sensibilidade	0.16239	0.8118	0.0006299	0.8789
Especificidade	0.94503	0.5749	1.0000000	0.8919
	Acurácia		0.62	

Fonte – Jailma Januário, 2021

O modelo de árvore de decisão apresentou uma boa acurácia, com o valor de 0.73, e a macro média de sensibilidade moderada com o valor de 0.60 demostrado na tabela 20, sendo esta a proporção geral do modelo para os verdadeiros positivos. Dentre os modelos, a árvore de decisão foi a que apresentou o melhor valor para a especificidade (0.89). A Tabela 14 ilustra a matriz de confusão gerada pela árvore de decisão. Neste modelo a classe que teve o maior número de instâncias corretas foi a classe F (Formado) com 37733 instâncias classificadas corretamente, seguida da classe D (desvinculado).

A Tabela 15 apresenta o valor das métricas de sensibilidade, especificidade para cada classe e a acurácia geral do modelo. Como podemos observar a sensibilidade demostrada pela árvore de decisão para a classe com maior representatividade, a classe D (desvinculado) é 0.99. Indicando uma alta proporção de verdadeiros positivos, já para a classe F (formado) o classificador apresenta uma sensibilidade baixa, com um valor de 0.21. Com relação à

Tabela 14 – Matriz de confusão Árvore de decisão

Classe prevista				
Classe real	T	D	TC	F
T	18368	11885	1126	22
D	17441	34892	1328	62
TC	227	240	683	8
F	339	608	38	37733

Fonte – Jailma Januário, 2021

especificidade o modelo apresenta uma proporção de 0.99 de verdadeiros negativos para a classe F (formado).

Tabela 15 – Medidas de desempenho por classe modelo Árvore de decisão

	T	**D**	**TC**	**F**
Sensibilidade	0.7326	0.9976	0.5050	0.215118
Especificidade	0.7566	0.9887	0.8529	0.996101
	Acurácia		0.73	

Fonte – Jailma Januário, 2021

O modelo de Regressão logística apresenta um valor para a média de acurácia de 0.54, demostrada na Tabela 20, pode-se observar também que o classificador demostrou uma proporção moderada para os verdadeiros positivos, apresentando um valor de 0.55 para a métrica de sensibilidade. A Tabela 16 apresenta os valores da matriz de confusão gerada pelo modelo de regressão logística. Observamos que a classe que o algoritmo mais acertou foi a F (formado) com 37428 instâncias do conjunto de teste, seguida pela classe D (desvinculado) com 23120 instâncias. Por outro lado, o quantitativo de instâncias classificadas corretamente como TC (transferência) foi muito baixo, apresentando um quantitativo alto de erro com relação a essa classe.

A Tabela 17 demostra as métricas utilizadas para a avaliação do modelo de regressão logística por classe. Pode-se observar que a métrica de sensibilidade para a classe D (desvinculado) é de 0.98. Isso indica que o modelo tem uma boa proporção de verdadeiros positivos, diferente da classe TC (transferência), onde a métrica de sensibilidade apresenta um valor de 0.15, indicando que o modelo não tem uma boa proporção de verdadeiros positivos. O classificador apresenta valores a partir de 0.73 para os verdadeiros negativos.

Tabela 16 – Matriz de confusão Regressão Logística

Classe prevista				
Classe real	T	D	TC	F
T	5574	4955	248	26
D	15595	23120	833	3
TC	13916	17572	1885	368
F	1290	1978	209	37428

Fonte – Jailma Januário, 2021

Tabela 17 – Medidas de desempenho por classe modelo de Regressão Logística

	T	**D**	**TC**	**F**
Sensibilidade	0.4855	0.9895	0.15324	0.59370
Especificidade	0.7876	0.9601	0.94100	0.73851
	Acurácia		0.54	

Fonte – Jailma Januário, 2021

O modelo de Redes Neurais Artificiais apresentou um valor de 0.67 de acurácia, 0.49 para a métrica de sensibilidade, demostrando uma baixa proporção para os verdadeiros positivos, e um valor de 0.86 para a métrica de especificidade demostrada na Tabela 20, indicando que este modelo tem um bom desempenho em classificar os falsos negativos. A Tabela 18 apresenta a matriz de confusão gerada a partir do modelo de Redes Neurais Artificiais. Pode-se observar que as classes que o algoritmo mais acertou foram as classes com maiores representatividades no conjunto de dados, a classe D (desvinculado) e F (formado), com 46242 e 37823 respectivamente. Por outro lado, o modelo não conseguiu classificar nenhuma instância da classe TC (transferência), que possui a menor representatividade no conjunto de dados.

A tabela 19 demostra os valores das métricas utilizadas para a avaliação do modelo, por cada classe do conjunto de dados. Pode-se observar que os valores de sensibilidade para as classes com maiores representatividades foram altos, apresentando os valores de 0.97 para a classe D (desvinculado) e 0.99 para a classe F (formado), ou seja, o algoritmo classifica bem os verdadeiros positivos. Com relação à Especificidade o modelo apresenta uma proporção mediana dos verdadeiros negativos para a classe D (desvinculado), porém para as outras classes o algoritmo apresenta valores altos para esta métrica.

Tabela 18 – Matriz de confusão Redes Neurais Artificiais

	Classe prevista			
Classe real	T	D	TC	F
T	890	927	48	0
D	35300	46242	2957	2
TC	0	0	0	0
F	185	456	170	37823

Fonte – Jailma Januário, 2021

Tabela 19 – Medidas de desempenho por classe modelo de Redes Neurais

	T	**D**	**TC**	**F**
Sensibilidade	0.02447	0.9710	0.0000	0.9999
Especificidade	0.98900	0.5055	1.0000	0.9907
	Acurácia		0.67	

Fonte – Jailma Januário, 2021

5.0.2 Comparação dos classificadores

Comparando os modelos a partir da média de acurácia e das macro-médias de sensibilidade e especificidade, pode-se observar na Tabela 20 que o modelo de árvore de decisão foi o que apresentou o melhor desempenho em todas as métricas, com destaque para a métrica de sensibilidade, pois para essa categoria de problemas esta é uma métrica muito utilizada para saber se o modelo é um bom classificador, visto que, apresenta uma boa identificação de verdadeiros positivos. Esta métrica é importante para este categoria de problema porque a não identificação do aluno que está em risco de evasão é mais prejudicial, que a identificação de um aluno que não está neste contexto. Dessa forma, o modelo deve ter alta sensibilidade, pois classificar alunos com risco de evasão como não tendo esse risco a instituição não terá como desenvolver ações para evitar que o aluno evada. No entanto, ter um modelo que classifica muitos falsos positivos podem elevar o custo e desperdiçar o tempo da gestão da instituição, com isso é necessário considerar a precisão do modelo, assim sendo também será avaliado para os classificadores a média harmônica entre a sensibilidade e a precisão, a métrica Fscore. Para esta métrica o melhor classificador continua sendo as árvores de decisão com o valor de 0.64, seguida pelo classificador Naive

Bayes com o valor de 0.55, com isso podemos afirmar que as árvores de decisão além de classificar bem os verdadeiros positivos, também pode ser utilizada para a tomada de decisão da gestão das instituições com relação às ações que podem desenvolver para evitar que o aluno abandone o curso.

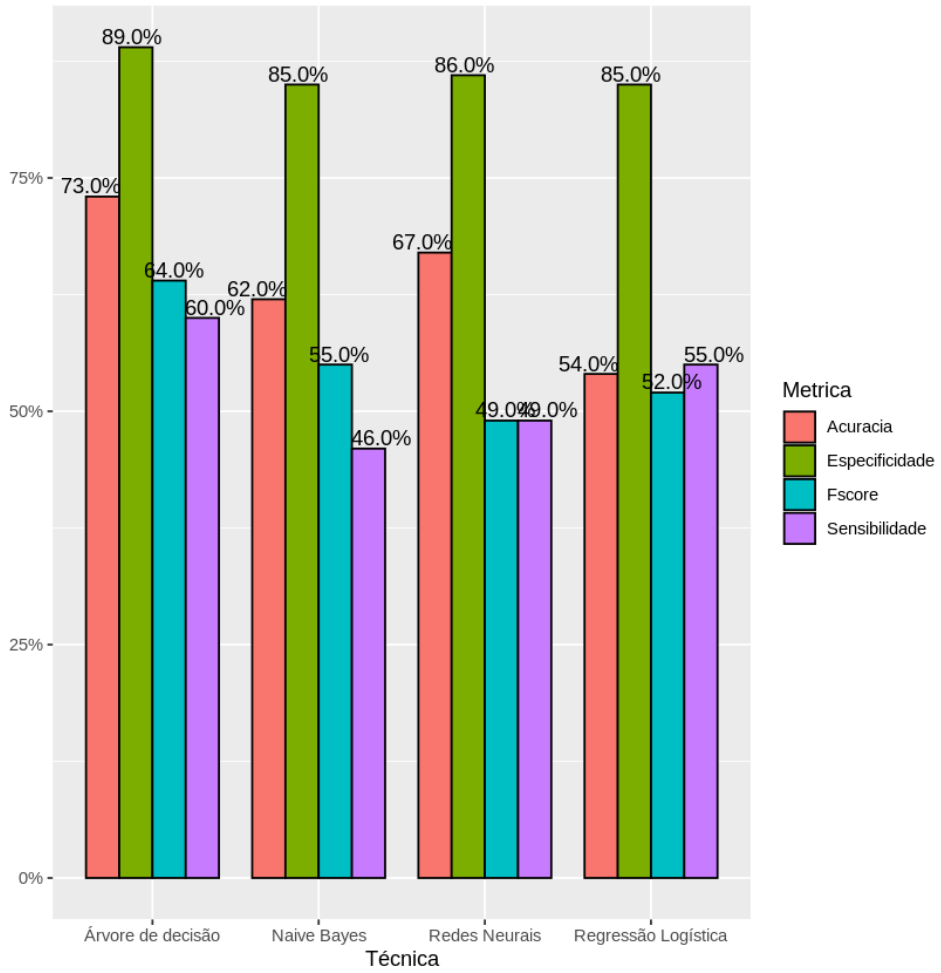
Tabela 20 – Valor de acurácia média e macro média para a sensibilidade e especificidade de cada modelo

	Acurácia	**Sensibilidade**	**Especificidade**	**FScore**
Naive bayes	0.62	0.46	0.85	0.55
Árvore de decisão	0.73	0.60	0.89	0.64
Redes neurais artificiais	0.67	0.49	0.86	0.49
Regressão logística	0.54	0.55	0.85	0.52

Fonte – Jailma Januário, 2021

A Figura 15 apresenta de forma gráfica o percentual da técnica por métrica de avaliação utilizada. A técnica de árvore de decisão escolhida como o melhor classificador para este problema está entre os melhores modelos para a classificação binária, sendo essa uma das principais motivações para a escolha dessa técnica em vários trabalhos, como pode ser visualizado na Seção 2. Este trabalho demostrou também que essa técnica alcança bons resultados para a classificação multiclasse com relação às métricas de acurácia, sensibilidade, especificidade e fscore. Pode-se destacar também que os demais classificadores obtiveram um bom resultado para esta categoria de problemas considerando especificidade e a métrica fscore, já para a métrica de sensibilidade apresentam uma proporção mediana de verdadeiros positivos.

Figura 15 – Valor das métricas por modelo



Fonte – Jailma Januário, 2021

5.0.3 Análise estatística

A análise estatística foi realizada aplicando o teste de Tukey para verificar se há diferenças estatísticas entre os modelos e entre as classes do conjunto de dados. Para a realização deste teste, em um primeiro momento é preciso fazer a análise de variância (ANOVA), para verificar se existe ou não diferenças entre as médias. Dessa forma, foram definidas as hipóteses com relação às técnicas e as classes do conjunto de dados. As hipóteses para as técnicas são:

- H_0: não existem diferenças significativas entre o desempenho dos classificadores para a métrica de acurácia;
- H_1: há ao menos um classificador com a métrica de acurácia que apresente diferença significativa.

As hipóteses para as classes são:

- H_0: não existem diferenças significativas entre as k classes para a métrica de acurácia;
- H_1: há ao menos uma classe com a métrica de acurácia diferente estatisticamente.

As informações geradas a partir da análise de variância podem ser visualizadas na Tabela 21, onde são descritos os graus de liberdade (GL), a soma de quadrados (SQ), o quadrado médio (QM), a estatística F e o valor-p. Para as técnicas e para as classes.

Tabela 21 – Análise de variância dos modelos e das classes por acurácia

	GL	**SQ**	**QM**	**F**	**Valor P**
Algoritmo	3	0.0204	0.00679	3.594	0.0591
Classe	3	0.4415	0.14716	77.911	9.27e-07
Residuos	9	0.0170	0.00189		

Fonte – Jailma Januário, 2021

Vemos um efeito significativo com relação à classe utilizada, com apenas um limiar de significância para as diferenças observadas entre os algoritmos. Isso significa que, geralmente, a classe contribuiu mais para as diferenças em acurácia do que o algoritmo em si. Dessa forma, a hipótese h_0 foi rejeitada com relação à classe, ou seja, ao menos um par de médias diferem entre si. Já com relação ao Algoritmo, não há evidências suficientes para afirmar se há diferença estatística entre as médias.

Vejamos agora uma comparação par a par, para essas duas variáveis. Para saber quais classes e algoritmos diferem entre si através da utilização do teste de Tukey com um nível de confiança de 95%. A Figura 16 apresenta os valores obtidos através da aplicação do teste para a variável algoritmo.

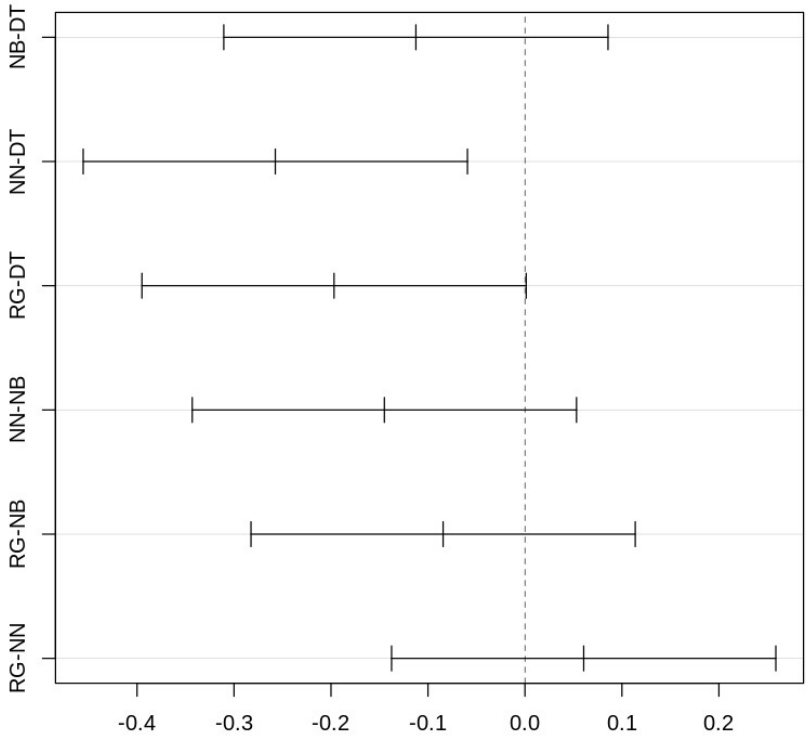
Figura 16 – Resultados do teste Tukey para as técnicas

	diff	lwr	upr	p adj
NB-DT	-0.09742090	-0.19335841	-0.001483389	0.0464966
NN-DT	-0.07124793	-0.16718544	0.024689586	0.1647130
RG-DT	-0.05938065	-0.15531816	0.036556861	0.2809073
NN-NB	0.02617297	-0.06976454	0.122110486	0.8288827
RG-NB	0.03804025	-0.05789726	0.133977761	0.6204166
RG-NN	0.01186728	-0.08407024	0.107804786	0.9792298

Fonte – Jailma Januário, 2021

Pode-se observar na figura17 que o teste post-hoc de Tukey, ajustado para múltiplas testagens, mostra haver diferença significativa, ainda que no limiar de significância, apenas entre Naive Bayes e Árvores de Decisão (NB-DT), apresentando um valor $p < 0,05$. As diferenças observadas entre os demais não foram significativas.

Figura 17 – Diferença entre as médias para as técnicas



Fonte – Jailma Januário, 2021

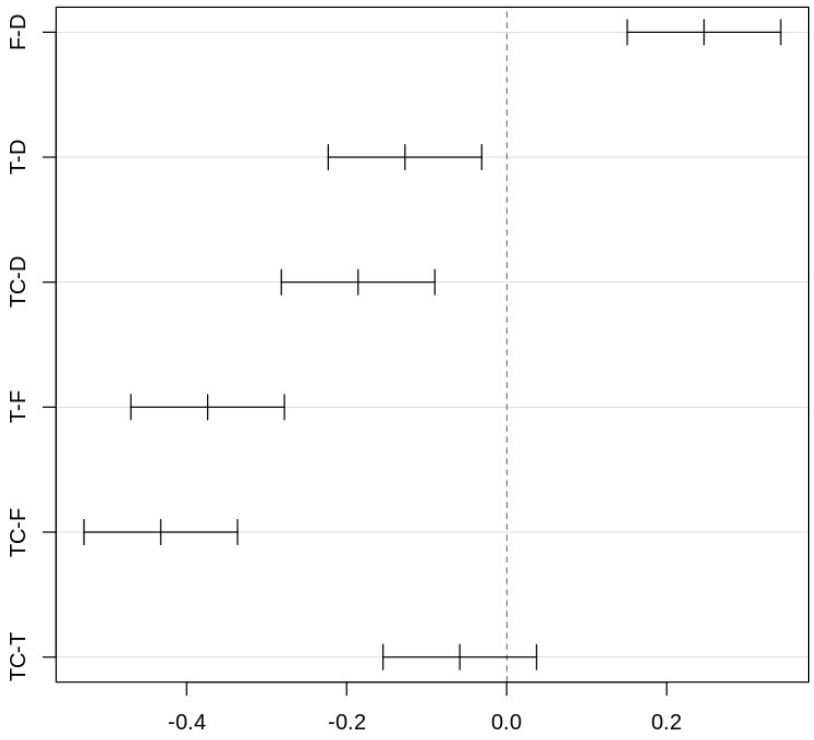
A Figura 18 apresenta o teste para a classe à qual cada dado pertence, como pode-se observar só não foram significativas as diferenças observadas entre as classes TC e T. As demais combinações resultaram em diferenças significativas, ilustradas na Figura 19.

Figura 18 – Resultados do teste Tukey para as classes

	diff	lwr	upr	p adj
F-D	0.2466275	0.1506900	0.34256499	0.0001034
T-D	-0.1271038	-0.2230413	-0.03116631	0.0111515
TC-D	-0.1857036	-0.2816411	-0.08976611	0.0008977
T-F	-0.3737313	-0.4696688	-0.27779379	0.0000033
TC-F	-0.4323311	-0.5282686	-0.33639359	0.0000010
TC-T	-0.0585998	-0.1545373	0.03733771	0.2904511

Fonte – Jailma Januário, 2021

Figura 19 – Diferença entre as médias para cada classe



Fonte – Jailma Januário, 2021

Pode-se observar que o teste foi sensível as diferentes categorias, dessa forma, podemos dizer que importou sim, a classe à qual o dado pertence, ou seja, a categoria de abandono, em vez do algoritmo utilizado (à exceção da diferença observada entre Naïve Bayes e Árvores de Decisão).

6 Considerações finais

A evasão de estudantes no ensino superior é um tema recorrente que vem sendo analisado e discutido por pesquisadores, professores e gestores das instituições de ensino tanto privada como pública, estas discussões também se estendem ao ensino presencial e a distância, cada modalidade com suas particularidades. Ao longo dos anos técnicas computacionais estão sendo estudadas e aplicadas para mitigar este problema, como o desenvolvimento de modelos de aprendizado de máquina e análise mais profunda dos dados educacionais. Dessa forma novas áreas de estudo estão surgindo neste campo, como a mineração de dados educacionais, que tem como objetivo o estudo e aplicação de métodos de mineração de dados na educação, tais informações são encontrados nas instituições e nos repositórios disponibilizados pelo governo. Possibilitando estudantes e pesquisadores realizarem estudos sobre a educação no país.

Dessa forma este trabalho tinha como questão norteadora da pesquisa analisar se é possível identificar alunos com vínculos reais no sistema de ensino superior, utilizando os dados disponíveis pelo INEP. Para isso foram estabelecidos os objetivos geral e específicos do trabalho, o qual teve como objetivo geral aplicar técnicas de aprendizado de máquia nos dados disponibilizados pelo INEP para classificar a situação de estudantes no ensino superior. E como objetivos específicos verificar se é possível classificar o aluno em diferentes categorias usando os dados de instituições públicas, Aplicar as técnicas mais comuns em aprendizado de máquina para a evasão de curso, e por fim apresentar o melhor modelo de classificação multiclasse.

Para alcançar esses objetivos foi realizada uma revisão sistemática da literatura, onde foi possível estabelecer as principais técnicas de aprendizado de máquina para este tipo de problema que são árvores de decisão, Naive Bayes, Redes Neurais Artificiais e regressão logística. Estabelecer quais eram os atributos mais importantes para utilizar em um algoritmo de classificação, os atributos de informações pessoal/social, informações demográficas e acadêmicas, estão entre os mais utilizados para criar um perfil de um aluno com tendência a evadir. Além disso também foi preciso definir as métricas de avaliação, acurácia, sensibilidade, especificidade e fscore para encontrar o melhor modelo. Definidos os pontos principais para a pesquisa, foi seguida uma metodologia que é bastante utilizada no mercado, a CRISP-DM com ela foi possível chegar aos resultados desta pesquisa.

Analisando os estudos que resultaram da revisão sistemática, foi possível perceber que os estudos se concentram em uma categoria de problema, os problemas binários (evadido ou não evadido), porém a jornada do aluno no ensino superior é complexa, e o aluno pode evadir de forma definitiva do curso, ou de forma temporária, além de poder realizar transferências de curso. Estas categorias mesmo classificadas como evasão são distintas e tem suas particularidades. Dessa forma o presente trabalho realizou uma classificação multiclasse considerando as diferentes categorias que o aluno pode se ser classificado.

A população da pesquisa foram alunos do ensino superior na modalidade presencial, os dados desses alunos estão disponíveis no site do INEP, referente ao ano de 2018. Foi realizado um estudo nessa base de dados para identificar características importantes no perfil do aluno que está prestes a evadir, dessa forma foram analisadas a distribuição das variáveis idade, sexo, cor, raça, formas de ingresso e se o aluno recebe algum auxílio ou não, estas informações são classificadas como importantes para identificar o perfil do alunos que evadem. A partir dessa seleção de variáveis e do pré-processamento, o conjunto de dados final contém 500.000 observações, após alguns ajuste nos dados, como retiradas de variáveis com valores faltantes, os dados foram aplicados aos algoritmos de classificação. Como resultados tem-se as árvores de decisão como o melhor classificador para problemas multiclasse analisados nesse trabalho, com base nas métricas de avaliação acurácia, sensibilidade, especificidade e F-score.

Foram analisadas as classes para verificar se existem variâncias entre as diferentes categorias ou não, para as classes, o teste de tukey apresenta que há diferenças estatísticas entre as médias, indicando que as categorias de evasão são diferentes entre si.

As limitações da pesquisa, se deram por conta da quantidade de dados que foram baixados do site do INEP, e de limitações de recursos computacionais para processar todos os dados. Dessa forma foi preciso diminuir o conjunto de dados para que a aplicação e análise das técnicas fossem viáveis no tempo da pesquisa.

Como trabalhos futuros espera-se aplicar outras técnicas de aprendizado de máquina, como métodos essembles para comparar como o melhor classificador deste trabalho. Além de adicionar mais dados sobre o histórico acadêmico dos alunos, pois além dos dados que foram utilizados nessa pesquisa, as informações acadêmicas também estão entre as mais utilizadas.

Referências

AHMED, S. A.; KHAN, S. I. A machine learning approach to predict the engineering students at risk of dropout and factors behind: Bangladesh perspective. In: *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. [S.l.: s.n.], 2019. p. 1–6. Citado 3 vezes nas páginas 12, 23 e 28.

ANDRADE, M. O. Os gêneros e a evasão no ensino superior: estudo de caso da faculdade governador ozanam coelho. *Revista Científica Fagoc Multidisciplinar*, 2016. Citado na página 55.

ASSIS, L. R. S. d. Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados. *Universidade de Brasília*, Rio de Janeiro, RJ, 2018. Citado 5 vezes nas páginas 27, 40, 47, 59 e 60.

BAKER, R. S. J. d.; ISOTANI, S.; CARVALHO, A. M. J. B. d. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 2, 2011. Citado 2 vezes nas páginas 13 e 31.

BARROS, P. H. F. d. Proposta e teste de um modelo de aprendizado para sistemas de cibersegurança. *Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas*, São José do Rio Preto, 2018. Citado na página 43.

BIOLCHINI J., M. P. G. N. A. C. C.; TRAVASSOS, G. H. Systematic review in software engineering. *Systems Engineering and Computer Science Department, UFRJ, Rio de Janeiro*, 2005. Citado na página 16.

BOLFARINE, H.; BUSSAB, W. O. Elementos de amostragem. *Universidade de São Paulo Instituto de Matemática e Estatística*, São Paulo, 2004. Citado na página 50.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. Crisp-dm. *SPSS*, 2000. Citado na página 45.

COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1, p. 1–29, 2013. Citado 2 vezes nas páginas 12 e 13.

CUNHA, J. P. Z. Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. *Universidade de São Paulo*, São Paulo, 2019. Citado na página 59.

DIGIAMPIETRI, L. A.; NAKANO, F.; LAURETTO, M. d. S. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. *Rev. Grad. USP*, v. 1, n. 1, 2016. Citado na página 14.

DONG, Y.; PENG, C.-Y. J. Métodos de dados ausentes com princípios para pesquisadores. *SpringerPlus*, 2013. Citado 2 vezes nas páginas 48 e 49.

GALAR, M.; FERNANDEZ, A.; BARRENECHEA, E.; BUSTINCE, H.; HERRERA, F. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, Elsevier, 2011. Citado na página 35.

GONÇALVEZ T.C. SILVA, J. C. C. O. A. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. Revista Brasileira de Computação Aplicada, 2018. Citado na página 14.

HUTAGAOL, N.; SUHARJITO. Predictive modelling of student dropout using ensemble classifier method in higher education. *Advances in Science, Technology and Engineering Systems*, v. 4, n. 4, p. 206–211, 2019. Citado na página 26.

KEMPER, L.; VORHOFF, G.; WIGGER, B. Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, v. 10, n. 1, p. 28–47, 2020. Cited By 1. Citado 2 vezes nas páginas 32 e 50.

KITCHENHAM, B. Procedures for performing systematic reviews. *NICTA Technical Report, Keele University Technical Report*, 2004. Citado na página 16.

LANES, M. d. A.; ALCâNTARA, C. d. S. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. *Simpósio Brasileiro de Informática na Educação*, Ceará, CE, 2018. Citado na página 13.

MARANHãO, J. D.; VERAS, R. M. O ensino noturno na universidade federal da bahia: percepções dos estudantes. *Ensaio: avaliação e políticas públicas em Educação*, Rio de Janeiro, 2017. Citado na página 54.

MAYRA, A.; MAURICIO, D. Factors to predict dropout at the universities: A case of study in ecuador. In: *2018 IEEE Global Engineering Education Conference (EDUCON)*. [S.l.: s.n.], 2018. p. 1238–1242. Citado 3 vezes nas páginas 13, 14 e 26.

MENDES, J. V. M. Avaliação das regionais de uma empresa de telecomunicações, através de análise de cluster. *Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Federal de Uberlândia, Faculdade de Matemática*, Uberlândia-MG, 2017. Citado na página 44.

MITCHELL, T. M. Machine learning. *McGraw-Hill*, 1997. Citado 6 vezes nas páginas 31, 34, 36, 37, 38 e 39.

NAGY, M.; MOLONTAY, R. Predicting dropout in higher education based on secondary school performance. In: *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*. [S.l.: s.n.], 2018. p. 000389–000394. Citado na página 23.

PARDO, T. A. S.; NUNES, M. d. G. V. Aprendizado bayesiano aplicado ao processamento de línguas naturais. *Núcleo Interinstitucional de Linguistica Computacional*, São Paulo, 2002. Citado na página 39.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 624. Citado na página 12.

PELLUCCI, P. R. S.; PAULA, R. R. d.; OLIVEIRA, S. W. B. de; LADEIRA, A. P. Utilização de técnicas de aprendizado de máquina no reconhecimento de entiades nomeadas no português. *Revista Exacta*, Belo Horizonte, 2011. Citado na página 32.

PEREIRA, T. R.; ZAMBRANO, C. J. Application of decision trees for detection of student dropout profiles. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2017. p. 528–531. Citado 2 vezes nas páginas 13 e 50.

PEREZ; CASTELLANOS, C.; CORREAL, D. Applying data mining techniques to predict student dropout: A case study. p. 1–6, 2018. Citado 4 vezes nas páginas 14, 21, 24 e 50.

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*, Rio de Janeiro, RJ, v. 26, n. 100, 2018. Citado na página 13.

REIS, E. A.; REIS, I. A. Análise descritiva de dados. *Universidade Federal de Minas Gerais*, Minas Gerais, 2002. Citado na página 50.

ROCHA, C.; ZELAYA, Y.; SáNCHEZ, D.; PéREZ, A. Prediction of university desertion through hybridization of classification algorithms. In: . [S.l.: s.n.], 2017. v. 2029, p. 215–222. Citado 2 vezes nas páginas 13 e 24.

RODRíGUEZ-MUñIZ, L.; BERNARDO, A.; ESTEBAN, M.; DíAZ, I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *PLoS ONE*, v. 14, n. 6, 2019. Citado na página 25.

SANTOS, V. C. Avaliação de um classificador multiclass com abordagem one-vs-one usando diferentes classificadores binário. *Universidade Federal de Pernambuco*, Pernambuco-PE, 2017. Citado 4 vezes nas páginas 34, 35, 38 e 40.

SILVA, A. C.; PINTO, R. C. *et al.* Mineração de dados do sistema acadêmico do instituto federal do sudeste de minas gerais-campus juiz de fora. *Seminários de Trabalho de Conclusão de Curso do Bacharelado em Sistemas de Informação*, v. 2, n. 1, 2018. Citado na página 12.

SILVA, F. C. d.; CABRAL, T. L. d. O.; PACHECO, A. S. V. Evasão em cursos de graduação: uma análise a partir do censo da educação superior brasileira. *XVI Coloquio Internacional de Gestion Universitaria- CIGU*, Arequipa, Peru, 2016. Citado 4 vezes nas páginas 30, 31, 52 e 54.

SILVA, L. A. d.; PERES, S. M.; BOSCARIOL, C. Utilização de técnicas de aprendizado de máquina no reconhecimento de entiades nomeadas no português. *Elsevier*, Rio de Janeiro, 2016. Citado 5 vezes nas páginas 32, 33, 34, 40 e 41.

SILVEIRA R.F. VICTORINO, M. d. C. H. M. L. M. Educational data mining: Analysis of drop out of engineering majors at the unb - brazil. *IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2019. Citado 3 vezes nas páginas 13, 14 e 26.

SIVAKUMAR, S.; VENKATARAMAN, S.; SELVARAJ, R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, v. 9, n. 4, p. 1–5, 2016. Citado na página 36.

SOUZA, A. M. d. Machine learning e a evasão escolar - análise preditiva no suporte à tomada de decisão. *Universidade FUMEC, Faculdade de Ciências Empresariais*, Belo Horizonte, 2020. Citado 4 vezes nas páginas 29, 31, 32 e 39.

SULTANA, S.; KHAN, S.; ABBAS, M. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education*, v. 54, n. 2, p. 105–118, 2017. Citado na página 25.

TEODORO, L. d. 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 – RBIE*, 2020. Citado 3 vezes nas páginas 48, 50 e 51.

ZEVIANI, W. M.; JúNIOR, P. J. R.; BONAT, W. H. Modelos de regressão não linear. *58º RBRAS e 15º SEAGRO*, Campina Grande - PB, 2013. Citado 2 vezes nas páginas 32 e 33.