UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Um processo orientado a dados para geração de modelo de predição de evasão escolar

Thiago Medeiros Barros

Orientador: Prof. Dr. Luiz Affonso Henderson Guedes de Oliveira

Co-orientador: Prof. Dr. Ivanovitch Medeiros Dantas da Silva

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Computação da UFRN (área de concentração: Engenharia de Computação) como parte dos requisitos para obtenção do título de Doutor em Ciências.

Número de ordem PPgEEC: 278
Natal, RN, outubro de 2020

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Barros, Thiago Medeiros.
Um processo orientado a dados para geração de modelo de predição de evasão escolar / Thiago Medeiros Barros. - Natal, 2020.
119 f.: il.

tese (doutorado) - Universidade Federal do Rio Grande do Norte, de Tecnologia, Pós-Graduação em Engenharia Elétrica e de Computação, Natal, RN, 2020.
Orientador: Prof. Dr. Luiz Affonso Henderson Guedes de Oliveira.
Coorientador: Prof. Dr. Ivanovitch Medeiros Dantas da Silva.

1. Mineração de dados educacionais - Tese. 2. Evasão escolar - Tese. 3. Modelo preditivo - Tese. 4. Classes desbalanceadas - Tese. I. Oliveira, Luiz Affonso Henderson Guedes de. II. Silva, Ivanovitch Medeiros Dantas da. III. Título.

RN/UF/BCZM CDU 004:37.012

Elaborado por Ana Cristina Cavalcanti Tinôco - CRB-15/262

Ao meu filho, Miguel, por me
inspirar a tentar ser um exemplo
todos os dias.

Agradecimentos

Gostaria de iniciar meus agradecimentos ao nosso sofrido povo brasileiro! São os impostos pagos por essa gente que proporcionou meu acesso ao ensino médio, curso técnico, graduação, mestrado e, neste momento, o doutorado de forma gratuita e com qualidade! Desse enorme povo, destaco especialmente todos os meus mestres e colegas hoje de profissão, os professores de instituições públicas que, em sua maioria, tanto fizeram e fazem por essa sociedade, apesar da falta de reconhecimento e ataques, principalmente nesse momento negacionista o qual estamos atravessando (que tenho fé que irá passar!). Em nome do meu orientador Luiz Affonso e Ivanovitch, obrigado!

Dessa enorme população brasileira que me ajudou, há alguns mais especiais nessa trajetória. Meu filho, Miguel Lima Medeiros, pelo simples fato de existir. Um colega, há algum tempo, falou "nós se tornamos de fato adultos quando precisamos ser exemplo para alguém". Você é esse alguém que me faz tentar ser um exemplo. Apesar do prazer intelectual que essa tese proporcionou, ela é fruto de inúmeros sacrifícios pessoais, principalmente de tempo com minha família.

Desses momentos de sacrifício e ausência, agradeço a minha esposa, Marilia Gabriela, por toda generosidade, amor e suporte para com nosso núcleo familiar. Não tenho dúvidas que parcela desse êxito é seu. Obrigado por me aguentar nos piores dias, e compartilhar os melhores também! Em seu nome e de Gracinha e Titinha, também agradeço à sua família, que me acolheu como um filho.

Aos meus genitores, José Barros e Maria de Fátima Medeiros, que são meus exemplos e heróis. Obrigado por todo sacrifício e superação das situações mais adversas, a fim de me permitir possibilidades que não tiveram. Ao meu pai, primeiro Doutor de toda família, você foi minha inspiração em concluir esse projeto acadêmico. À minha mãe, que deixou de trabalhar e estudar por mim, essa tese também é sua.

Para minha irmã, Thalita Medeiros Barros, e meu cunhado Pedrinho. Obrigado pela enorme ajuda nesse período e suporte na criação de Miguel.

Aos meus avôs e avós, Severino Emídio de Medeiros, Alice Edite de Medeiros, Tutu Pereira, Antônio Barros, Rita Pereira, por cuidar, criar e educar meus pais. Certamente não estaria aqui sem vocês.

Aos maravilhosos tias e tios, agradeço em nome de Aninha, Vera, Ramos e Lúcia, pelo que vocês fazem por nossa família. Em nome de vocês, agradeço a todos os familiares.

Aos meus queridos vizinhos e amigos, Raffael, Michele e João Gabriel, a parceria, comida, godeladas, principalmente nesse momento tão difícil que foi a pandemia para nossas crianças.

Aos colegas do IFRN, agradeço em nome do nosso saudoso Alex Oliveira (*in memorian*) pelo apoio que essa instituição possibilitou, seja como aluno, seja como servidor.

Aos meus nadadores em mar aberto e corredores de asfalto que sempre me fazem lembrar a importância de estar aqui! Nesse momento! Em nome de César Careca e Paulo Bob, eu agradeço.

Aos amigos que moram longe, mas a amizade nunca morre, em nome de Rainery e Renata, agradeço sempre o apoio e incentivo.

Aos amigos que quase nunca vejo e sempre faz questão de lembrar que preciso aparecer, em nome de Ronkaly, agradeço a amizade.

Aos meus eternos bolsistas, os arcanjos Rafael e Gabriel, por serem meus braços em tantos projetos!

Por fim, a todos que divulgaram o seu conhecimento e contribuíram de alguma forma para a construção desta tese!

"Números, mas não só números. O mundo não pode ser compreendido sem números e não pode ser compreendido só com números. Aprecie os dados numéricos por aquilo que dizem sobre vidas reais. ". ROSLING, Hans. Factfulness. Flammarion, 2019.

Resumo

A evasão escolar, também conhecida como abandono escolar, é um problema extremamente complexo, pois envolve não apenas uma variedade de perspectivas, mas também uma variedade de diferentes tipos de comportamento de abandono. Historicamente, os modelos de evasão escolar mais citados tiveram sua origem na educação, entretanto a emergente área de Ciência de Dados aplicada na Educação é capaz de desenvolver novos modelos preditivos, com resultados geralmente melhores quando comparados com os métodos estatísticos tradicionais. O principal objetivo dessa tese é a proposição de um processo para geração de um modelo preditivo de evasão escolar baseada em Ciências de Dados. Para tal, uma sequência de etapas é definida, a fim de modelar um fluxo de informação, desde a definição do problema até a geração de informação útil a gestores e professores. As etapas são compostas por: "Entender o Problema", "Entender os Dados", "Engenharia de Atributos", "Seleção de Atributos", "Balanceamento de Dados", "Modelos", "Avaliação"e "Interpretação". A contribuição da proposta se encontra na indicação de quais técnicas e algoritmos devem ser empregados em cada etapa do processo apresentado, e na exposição de que o fenômeno de evasão escolar deve ser abordado como um problema de classes desbalanceadas, a qual deve utilizar-se de ferramentas e métricas apropriadas, a fim de gerar um modelo de predição robusto e de fácil interpretação. O processo proposto foi validado sobre dados educacionais, socioeconômicos e demográficos de alunos de cursos integrados do Instituto Federal do Rio Grande do Norte (IFRN).

Palavras-chave: mineração de dados educacionais, evasão, modelo preditivo, classes desbalanceadas.

Abstract

School dropout is an extremely complex problem, as it involves not only a variety of perspectives, but also a variety of different types of dropout behavior. Historically, the most cited school dropout models had their origin in education, however the emerging area of Data Science applied in Education is capable of developing new predictive models, with generally better results when compared to the most used traditional statistical methods. The main objective of this thesis is the proposition of a process for the generation of a predictive school dropout model based on Data Science. To this end, a sequence of steps is defined in order to model an information flow from problem definition to generation of useful information for managers and teachers. The steps consist of: Understanding the Problem, Understanding the Data, Feature Engineering, Feature Selection, Data Balancing, Models, Evaluation and Interpretation. The proposal's contribution is found in the indication of which techniques and algorithms should be used in each phase of knowledge discovery, and show that the phenomenon of school dropout must be addressed as a problem of imbalanced classes, and should be approached with appropriate tools and metrics, in order to generate a robust and easy to interpret prediction model. The proposed process was validated on educational and socioeconomic data of students Federal Institute of Rio Grande do Norte (IFRN).

Keywords: educational data mining, dropout, predictive model, imbalanced classes.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de símbolos e abreviaturas

AUC	Area under curve
DT	Decision Tree
EDA	Exploratory Data Analysis
EDM	Educational Data Mining
EWS	Early Warning System
FN	False Negative
FP	False Positive
FPR	False positives rates
$H0$	Hipótese nula
KNN	K-Nearest Neighbours
LA	Learning Analytics
LR	Logistic Regression
MCC	Matthews correlation coefficient
NB	Naive Bayes
RF	Random Forest
$RFECV$	Recursive Feature Elimination and Cross-Validated Selection
ROC	Receiver Operating Characteristic
$SGDC$	Stochastic Gradient Descent
$SMOTE$	Synthetic Minority Oversampling Technique
SVM	Support vector machine
TN	True Negative
TP	True Positive

TPR	True positives rate
UAR	Unweighted Average Recall
AB	Acurácia Balanceada
AC	Análise de Correspondência
AM	Aprendizado de Máquina
AVA	Ambiente Virtual de Aprendizagem
EaD	Educação a Distância
IFRN	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
IPEA	Instituto de Pesquisa Econômica Aplicada
RN	Rio Grande do Norte
RNA	Redes Neurais Artificiais
RSL	Revisão Sistemática da Literatura
V/F	Verdadeiro ou Falso
VIF	Variância do fator de influência

Capítulo 1

Introdução

A evasão escolar, também conhecida como abandono escolar, é um problema complexo, pois envolve não apenas uma variedade de perspectivas, mas também de diferentes tipos de comportamento de abandono (Tinto 1982). É importante ressaltar que a definição de evasão escolar depende da perspectiva adotada. O Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)[1], por exemplo, no Art. 227 do seu documento de organização didática, caracteriza evasão escolar como: "Terá matrícula cancelada por EVASÃO o estudante que não efetuar a renovação de matrícula, em qualquer período do curso". Logo, podemos entender que a evasão escolar para o IFRN ocorre quando um aluno deixa de frequentar a escola.

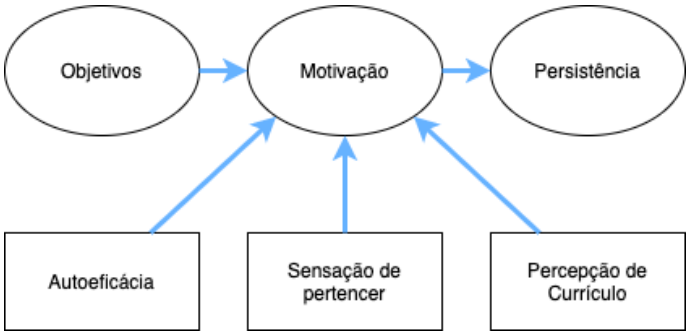
A evasão escolar representa oportunidades de mudança de vida desperdiçadas, menos mão-de-obra qualificada no mercado e menor chance de mobilidade social (Barros 2017), principalmente em um país com índices de desigualdades sociais como o Brasil. A seguir, vemos alguns números para ilustrar a evasão no país: em 2010, 11,4% dos alunos abandonaram o curso para o qual foram admitidos. Já em 2014, esse número chegou a 49% (BRASIL 2016*a*), em um país que apresenta o percentual de 75% de jovens de 20 a 24 anos de idade que não estudam, sendo o maior índice no mundo entre os países pesquisados pelo relatório *Education at a Glance* (BRASIL 2016*b*). Além disso, estima-se que 7 bilhões de reais por ano seja o valor investido em 1,9 milhão de jovens de 15 a 17, que acabam abandonando o ensino médio antes do final do ano ou são reprovados ao final deste (Barros 2017), valor equivalente ao custeio de todos Institutos e Universidades Federais do país no ano de 2017 (BRASIL 2018). No contexto de uma nação com altas taxas de homicídios como o Brasil, é importante destacar o estudo do Instituto de Pesquisa Econômica Aplicada (IPEA), o qual aponta que, a cada adicional de 1% na taxa de atendimento escolar (que corresponde ao número de matriculados em relação à população de sua respectiva faixa etária) de jovens entre 15 e 17 anos, os homicídios caem 1,9% (Cerqueira & Moura 2019). Portanto, é evidente a necessidade do desenvolvimento de ferramentas que possam ser utilizadas para minimizar o baixo desempenho e maximizar a persistência do aluno no ambiente escolar. Por isso, há grande interesse na comunidade científica por modelos conceituais que possam descrever adequadamente a evasão escolar (Tinto 1982), pois eles servem de base para melhor compreender, caracterizar, explicar e até prever esse importante fenômeno.

[1] Serão utizados dados escolares do IFRN no Estudo de Caso pra validação da proposta desta tese

Historicamente, os modelos sobre evasão escolar mais adotados na literatura têm origem na área de educação (Tinto 1975) e são baseados quase que completamente em conceitos subjetivos associados com aprendizagem e motivação, exceto por algumas medidas de estatísticas descritivas básicas. Assim, esses modelos são de natureza mais qualitativa. Por outro lado, a emergente área de Ciência de Dados aplicada a dados educacionais possibilitou o desenvolvimento de modelos quantitativos fortemente baseados em dados para caracterização do fenômeno da evasão escolar, com capacidade tanto de caracterizar os fatores que contribuem substancialmente para o fenômeno quanto de servir de ferramenta de predição (Thammasiri et al. 2014).

Dentre os modelos conceituais sobre evasão escolar, se destaca o proposto por Tinto (Tinto 1975). Essa abordagem afirma que a decisão dos estudantes de persistir ou evadir dos estudos está fortemente relacionada ao grau de integração acadêmica e integração social na instituição educacional. Dessa forma, o autor define que a integração social inclui muitos aspectos do cotidiano dos alunos, como amizades, apoio familiar e sentimento de satisfação, e a integração acadêmica como o conjunto de regras, normas e expectativas acadêmicas. No seu trabalho mais recente, Tinto propõe que a persistência é fortemente relacionada com a motivação do estudante, que por sua vez está relacionada com: Objetivos, autoeficácia, sensação de pertencer e percepção de currículo (Figura 1.1). De forma sucinta, "Objetivos"representam as intenções do aluno em concluir aquele curso naquela instituição, "Autoeficácia"é relacionada ao sentimento da capacidade do aluno em concluir os estudos, a "Sensação de Pertencer"representa a integração social do aluno na instituição e "Percepção de Currículo"é vinculada à relevância daquele curso para o aluno e da integração acadêmica na instituição (Tinto 2017).

Figura 1.1: Um modelo de motivação e persistência do aluno.



Fonte: Adaptada de Tinto (2017)

Entretanto, o modelo conceitual de Tinto parece ter duas grandes desvantagens: falta de generalização e alto custo envolvidos na realização de pesquisas em grande escala (Delen et al. 2019).

Outros modelos conceituais sobre evasão escolar também bastante citados na literatura são a Teoria do envolvimento de Astin (Astin 1999) e o Modelo de atrito dos alunos de Beans (Bean 1982). O trabalho de Bean forneceu uma explicação alternativa, em que a decisão de evadir do estudante é motivada por sua experiência com diferentes aspectos

da instituição, como qualidade institucional, professores e amigos. Da mesma forma, a teoria de Astin afirma que os estudantes são mais afetados por três tipos de envolvimento: com a instituição de ensino, com os professores e com o grupos de colegas. Quanto maior o esforço que os alunos investem nesses envolvimentos, maior a probabilidade de persistirem. Embora as três teorias tenham estruturas conceituais diferentes, o consenso é de que a integração social e o relacionamento com os colegas são um indicador importante para persistência deles.

Por outro lado, há na literatura diversos trabalhos baseados em Ciência de Dados e Inteligência Artificial, geralmente focados nos modelos de aprendizagem para predição de performance ou evasão do aluno. Nenhum deles enfatiza uma visão mais genérica como um processo sistemático para geração de modelos de predição de evasão escolar. Essas pesquisas são capazes de lidar com grande quantidade de dados e outras restrições como valores ausentes, dependência, correlação, desbalanceamento de dados, alta dimensionalidade, normalidade e relações não-lineares, produzindo melhores precisões na predição. Entretanto, é observada a falta de consenso sobre quais técnicas e algoritmos devem ser adotados nas diversas atividades associadas ao problema de predição de evasão baseada em dados. Com isso, para cada caso de estudo, há a necessidade de se refazer todo o procedimento de modelagem. Assim, torna-se clara a necessidade de se estabelecer um processo padronizado para modelar o fluxo, indo desde a obtenção de informação sobre evasão escolar (que inclua atividades da captura de dados brutos) até a disponibilização de informação em formato útil a gestores e professores.

É importante ressaltar que, ao criar modelos baseados em Ciências de Dados e Inteligência Artificial, algumas perguntas frequentes são: "Como posso confiar nesse seu modelo?", "Como seu modelo toma suas decisões?"(Ribeiro et al. 2016, Mori & Uchihira 2018, Johansson et al. 2011). Logo, é necessário também observar o aspecto qualitativo do modelo, a partir da sua interpretação, a fim de identificar e evitar possíveis vieses contidos nos dados (até mesmo preconceituosos), tornando os modelos mais justos (Corbett-Davies & Goel 2018). Por isso, nesta tese defendemos que as abordagens qualitativas e quantitativas de modelagem do problema de evasão escolar não são concorrentes entre si, mas complementares. Assumimos, assim, que se deva utilizar as vantagens de ambas as abordagens para se construir a melhor solução para modelar o possível problema da evasão escolar.

1.1 Motivação da tese

Defendemos nesta tese que a emergente área de Ciência de Dados aplicada na Educação é capaz de desenvolver melhores modelos de evasão escolar, tanto para caracterização dos fatores relevantes quanto para predição do fenômeno quando comparada com os métodos estatísticos tradicionais, por apresentar menores restrições (por exemplo, normalidade, independência e colinearidade) (Thammasiri et al. 2014). A importância de se ter bons modelos preditivos diz respeito ao fato de que, quanto mais cedo se detectar uma possível desistência ou insucesso na atividade acadêmica do aluno, maior a chance de uma intervenção evitar a continuação de seu baixo desempenho, ou até a sua evasão escolar (Delen 2011, Thammasiri et al. 2014, Burgos et al. 2018, Nelson et al. 2012, Jayaprakash

et al. 2014, Romero & Ventura 2019*b*).

É importante enfatizar a característica do problema da evasão escolar como um fenômeno contextualizado e complexo em cada instituição de educação (Monllaó Olivé et al. 2019). Essas características vão ao encontro ao teorema *No silver bullet model*, em que não há modelo perfeito para qualquer tipo de dados (Wolpert 2002). Portanto, acreditamos que definir um processo sistemático para gerar o modelo, mais do que o desenvolvimento de uma máquina de aprendizagem de propósito geral complexa, é a forma mais adequada para contemplar as especificidades do problema de evasão para cada instituição de educação em particular.

1.2 Objetivos da tese

O principal objetivo desta tese é a proposição de um processo sistemático para geração de modelos baseados em dados, a fim de prever o fenômeno da evasão escolar. Assim, em certa medida, esse processo sistemático pode ser considerado como um metamodelo de evasão escolar. Para tal, é definida uma sequência de etapas, com o finalidade de modelar um fluxo de informação desde a definição do problema até a geração de informação útil a gestores e professores. As etapas são compostas por: "Entender o Problema", "Entender os Dados", "Engenharia de Atributos"(mais conhecido por *Feature Engineering* em inglês), "Seleção de Atributos"(mais conhecido por *Feature Selection* em inglês), "Balanceamento de Dados", "Modelos", "Avaliação"e "Interpretação". Como objetivos secundários, as etapas serão norteadas para buscar a obtenção de modelos simples (Shavlik et al. 1990), de fácil compreensão ao humano e ênfase em identificar o aluno que potencialmente possa evadir à identificação do aluno persistente.

A partir da definição do processo sistemático, também temos como objetivo verificar a hipótese: "o modelo baseado em dados de evasão escolar gerado tem um desempenho melhor frente a suposição que o aluno irá evadir caso: (I) seu desempenho médio e (II) sua frequência média estejam inferiores ao desempenho mínimo estabelecido pela organização didática".

1.3 Contribuição da tese

A principal contribuição conceitual desta tese é a proposição de um processo sistemático para obtenção de modelos baseados em dados para evasão escolar. Em decorrência disto, tem-se as seguintes contribuições na área de modelagem baseada em dados educacionais: avaliação de técnicas e algoritmos empregados em cada fase da descoberta do conhecimento; caracterização do fenômeno de evasão escolar como um problema de classificação de dados com classes desbalanceadas; e avaliação de ferramentas e métricas de desempenho a fim de se obter modelos robustos e de fácil interpretação (Barros, Souza Neto, Silva & Guedes 2019). Para isso, são investigadas técnicas e algoritmos de pré-processamento de dados, análise estatística (Barros 2019), aprendizagem de máquinas e visualização científica de dados (Barros 2017, Barros 2018).

O processo sistemático proposto aqui será validado sobre dados educacionais, socioeconômicos e demográficos de alunos de cursos integrados do Instituto Federal do Rio Grande do Norte (IFRN). Essa base de dados foi escolhida para análise por sua qualidade de informação e por ser representativa no critério da diversidade da caracterização social de alunos por todo estado. É importante enfatizar que a construção desta base, a partir dos dados extraídos do sistema acadêmico do IFRN, também é fruto deste trabalho e poderá ser utilizada por outros estudos.

Por fim, defendemos que as contribuições descritas nesta tese podem servir como suporte para instituições de educação, órgãos de controle educacional, como o Ministério da Educação (MEC), secretarias e conselhos educacionais para a tomada de decisão a cerca do relevante problema da evasão escolar.

1.4 Organização da tese

Neste primeiro capítulo, foi apresentada uma introdução da tese, descrevendo o contexto, as motivações, os objetivos e as contribuições do trabalho. O restante do documento é dividido por mais cinco capítulos:

- No capítulo 2, a área de Ciências de Dados é apresentada e as técnicas e modelos utilizados em cada uma das fases do Processo proposto são descritos;
- No capítulo 3, é realizada uma revisão sistemática da literatura, a fim de posicionar este trabalho no estado da arte de modelos de predição de evasão escolar baseada em dados;
- No capítulo 4, o processo proposto para análise sistemática de dados educacionais é apresentado, indicando-se suas fases, técnicas e algoritmos a serem empregados;
- No capítulo 5, é feita a apresentação das configurações experimentais para propósito de validação da proposta. Nesse intuito, o estudo de caso, a base de dados utilizada para validação do processo proposto, o ambiente de desenvolvimento e os resultados obtidos em cada etapa do processo proposto são apresentados.
- No capítulo 6, são expostas as conclusões e os trabalhos que serão desenvolvidos posteriormente.

1.5 Trabalhos publicados

O primeiro trabalho publicado desta tese foi a partir da criação e análise exploratória de dados sobre a base de dados brutos extraídos do sistema de gestão acadêmica do IFRN. No trabalho (Barros et al. 2017), foi apresentado o potencial do uso de técnicas de visualização de dados para detectar padrões que não seriam possíveis analisando apenas métricas estatísticas mais tradicionais como médias, variâncias e gráficos como o boxplot. Por exemplo, o uso do gráfico violino permitiu a detecção de padrão bimodal corroborados pelo gráfico Q-Q, o que não seria possível se fosse utilizado apenas o boxplot. Uma vez que foi detectado o aparecimento de mais de um grupo nos dados, a técnica de k-means se mostrou eficaz em separar esses grupos, o que foi corroborado pela técnica de análise de correspondência

A partir das primeiras descobertas de conhecimento do trabalho anterior, percebeu-se a predominância de dados categóricos na base, principalmente aqueles de natureza socioeconômica e demográfica. Nos trabalhos (Barros et al. 2018) e (Barros, Silva & Guedes 2019), há a proposição de uma sistematização para identificação do perfil do aluno evadido a partir da técnica análise de correspondência, devido à natureza categórica de boa parte dos dados sobre os estudantes. Logo, a determinação do grau de independência entre variáveis via uso de técnica de análise de correspondência se mostrou mais adequada do que a tradicional análise de correlação de Pearson.

Por fim, após o vasto entendimento dos dados proporcionados pelos trabalhos anteriores, verificamos que a evasão escolar deveria ser abordada como um problema de classes desbalanceadas. Assim, iniciou-se o processo de pré-processamento dos dados, aplicação de modelos de aprendizagem e avaliação desses modelos. Essas atividades foram publicadas em (Barros, Souza Neto, Silva & Guedes 2019), em que concluímos que as métricas de precisão, *Recall* e F1 não conseguiram detectar a grande quantidade de erros da classe minoritária (o aluno evadido) quando os dados estavam desequilibrados. No entanto, verificamos que as métricas *G-mean* e UAR foram capazes de capturar o erro de classe minoritária, mostrando que essas são mais indicadas para avaliação de problemas com classes desbalanceadas. Concluímos também que o uso da técnica de balanceamento de dados, antes do treinamento do modelo preditivo, promove uma melhora significativa nos resultados quando mensurados pelas métricas *G-mean* e UAR.

Capítulo 2

Fundamentação teórica - Ciência de Dados

Nesse capítulo, serão apresentados os conceitos de Ciência de Dados aplicados na área de Educação, além dos algoritmos e técnicas que servirão de base para o processo sistemático para geração de modelos preditivos de evasão escolar.

Se a Ciência é uma atitude cética em relação aos fatos, criando modelos que tentam explicar a realidade ou testando hipóteses com uma metodologia rígida para outras pessoas que tentam reproduzir e falsificar (Popper 2004), a Ciência de Dados é definida como uma metodologia pela qual modelos e hipóteses podem ser inferidos a partir de dados (Igual & Seguí 2017). A Ciência é um ciclo de conjecturas e refutações, em que se aumenta a confiança em uma hipótese à medida que as provas se acumulam (Pinker & Motta 2018). Portanto, realizar a Ciência de Dados significa produzir conjecturas e refutações a partir dos dados, a fim de serem utilizadas como base para tomada de decisões. Logo, a representação de ambientes complexos a partir de dados abre a possibilidade de aplicar uma ampla gama do conhecimento científico que temos com o intuito de inferir modelos a partir deles (Igual & Seguí 2017).

Em geral, a Ciência de Dados adota quatro estratégias diferentes para exploração e análise dos resultados:

- **Examinando a realidade**: Os dados podem ser coletados por métodos passivos ou ativos. Neste último caso, os dados representam a resposta do mundo às nossas ações. A análise dessas respostas pode ser extremamente valiosas quando se trata de tomar decisões sobre nossas ações. Um dos melhores exemplos dessa estratégia é o uso de teste A / B para desenvolvimento da Web: qual é o melhor tamanho e cor do botão? A melhor resposta só pode ser encontrada examinando o mundo.
- **Descobertas de padrões**: Os problemas com dados podem ser analisados automaticamente para descobrir padrões e agrupamentos naturais. A partir do uso de algoritmos implementados em computadores, essa atividade se tornou extremamente eficaz e eficiente, principalmente quando comparada com a produtividade humana nesse tipo de análise. Um exemplo de aplicação no contexto educacional é a criação de perfis de alunos a partir de seus desempenhos em determinadas disciplinas.
- **Predição de eventos futuros**: Desde os primórdios da estatística, uma das questões científicas mais importantes tem sido como construir modelos robustos a partir de

dados que sejam capazes de prever novos eventos. Naturalmente, não é possível prever o futuro em qualquer ambiente e sempre haverá eventos imprevisíveis. Um exemplo de aplicação é o uso de modelos preditivos no contexto educacional para previsão de evasão e de desempenho do aluno.

- **Entendendo as pessoas e o mundo**: Este é um objetivo que, no momento, está além do escopo da maioria das empresas e pessoas. Porém, grandes instituições e governos estão investindo quantias consideráveis de recursos financeiros e humanos em áreas de pesquisa associadas a resolução desse problema, como compreensão da linguagem natural, visão computacional, psicologia e neurociência. A compreensão científica dessas áreas é importante para a Ciência de Dados porque, no final, para tomar decisões ideais, é necessário conhecer os processos reais que orientam as decisões e o comportamento das pessoas. Um exemplo de aplicação é entender quais atributos podem levar um aluno à depressão ou diminuir sua motivação nos estudos.

Como já descrito no capítulo de introdução, nesta tese propomos um processo sistemático para criação de modelos de predição de evasão escolar baseados em dados, ao invés de se propor um modelo de propósito geral. Nossa decisão se deve ao fato que o problema citado é um fenômeno contextualizado e complexo mesmo dentro de uma instituição de educação, quiçá entre instituições. Alguns fatores que corroboram com nossa decisão são: (a) os valores dos atributos (variáveis independentes do modelo) podem estar em diferentes faixas de valores nas diferentes bases de dados educacionais, dificultando a padronização e comparação entre essas bases; (b) a carga de trabalho entre os cursos são diferentes, logo os modelos preditivos precisam levar em consideração a dificuldade dos cursos para serem concluídos com êxito; (c) modalidade do curso, uma vez que os cursos totalmente *online* têm todas as atividades dos alunos registradas via infraestrutura computacional e em rede, diferentes dos cursos híbridos e presenciais (Monllaó Olivé et al. 2019). Esses fatos confirmam a máxima *No Silver Bullet*, em que não há modelo perfeito para qualquer tipo de dados (Wolpert 2002).

Uma das possíveis aplicações para esse processo sistemático proposto nesta tese seria a concepção de um tipo "Sistema de Aviso Prévio"(em inglês, *Early Warning System* - EWS), que se caracteriza por ser um sistema projetado para a prevenção de um problema antes que ele se torne um perigo real. No contexto educacional, um EWS acadêmico consiste em um conjunto de procedimentos e instrumentos para a detecção precoce de estudantes em risco de baixo desempenho ou evasão, além de intervenções apropriadas para mantê-los na instituição de educação (Romero & Ventura 2019*a*).

Assim, para embasar a nossa proposta, nas próximas seções serão apresentados as técnicas e os algoritmos utilizados nas diversas etapas do processo sistemático proposto nesta tese. As seções serão dividas em: (I) Análise exploratória de dados, que engloba um conjunto de técnicas para visualização de dados e extrair informações estatísticas iniciais das distribuições das variáveis, para auxiliar na toma de decisão nas próximas etapas; (II) Pré-processamento, o qual, a partir da compreensão inicial dos dados da etapa anterior, permite saber quais técnicas podem ser utilizadas para manipular diferentes tipos de dados e gerar uma nova base mais adequada para os algoritmos de aprendizagem na próxima etapa; (III) Modelos baseados em aprendizagem de máquina,o qual, a partir da base de

dados produzida pela etapa anterior, nos revela quais as vantagens e desvantagens dos diversos algoritmos de aprendizagem para gerar um modelo de predição; (IV) Avaliação de modelos de aprendizagem supervisionada, que nos informa as métricas mais interessantes para avaliar os modelos gerados; e (V) e Interpretação de modelos, que apresenta técnicas para auxiliar no entendimento pelos humanos dos modelos gerados.

2.1 Análise exploratória de dados

A análise exploratória de dados (em inglês, *Exploratory Data Analysis* - EDA) é uma maneira de fazer avaliações iniciais sobre a distribuição populacional das variáveis usando os dados das amostras observadas. Um dos principais objetivos da EDA é visualizar e resumir a informações estatísticas, permitindo-nos fazer suposições iniciais sobre a distribuição da população (Igual & Seguí 2017), além de maximizar a percepção de um conjunto de dados, descobrir estruturas subjacentes, extrair variáveis importantes, testar premissas, detectar *outliers* e anomalias (NIST/U.S.A 2013).

Para alcançar esses objetivos, algumas análises comuns sobre os dados são:

1. Dimensionalidade da base;
2. Dados ausentes;
3. Proporção entre as classes (Balanceamento);
4. Correlação (numéricos) e dependência (categóricos) entre atributos;
5. Tipo dos atributos;
6. Visualização das distribuição dos dados;
7. Estatística descritiva (média, mediana, mínimo, máximo, percentil);
8. Visualização dos dados em baixa dimensão (PCA, t-SNE);
9. Visualização das relação entre atributos.

2.2 Visualização científica de dados

Técnicas avançadas de visualização científica de dados deixaram de ser apenas assessórias para se tornarem essenciais na interpretação dos resultados, uma vez que a análise destes dependem fortemente de como os dados são apresentados (McCandless 2014). Em muitos casos, o emprego de técnicas de visualização de dados adequadas possibilita inferências complexas por parte dos usuários, mesmo utilizando-se apenas técnicas básicas de estatística. É importante ressaltar que técnicas de visualização não são substitutas ou concorrentes das técnicas estatísticas e de aprendizagem de máquina (Seção 2.4), mas sim aliadas que visam potencializar as análises sobre os dados.

A seguir, serão descritas algumas ferramentas utilizadas para visualização de estatísticas sobre os dados.

2.2.1 Gráfico violino

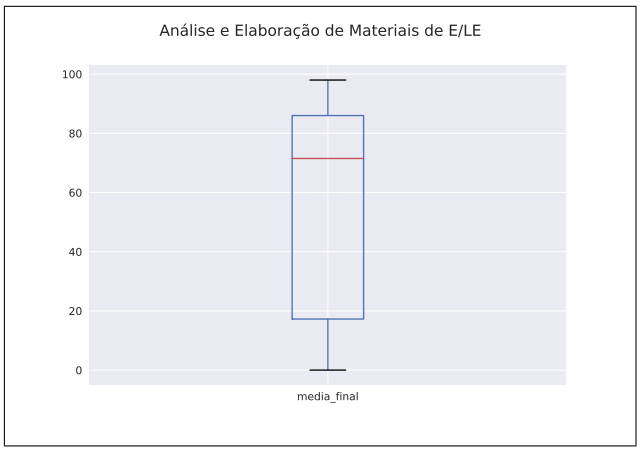
O gráfico tipo violino é a combinação do gráfico *boxplot* e a linha de densidade no mesmo diagrama (Hintze & Nelson 1998). O gráfico *boxplot* permite capturar caracte-

rísticas importantes dos dados em sua visualização, como a média, o espalhamento, a assimetria e os *outliers*. Essas características são identificadas a partir da indicação no gráfico do limite inferior, primeiro quartil, mediana, terceiro quartil, limite superior.

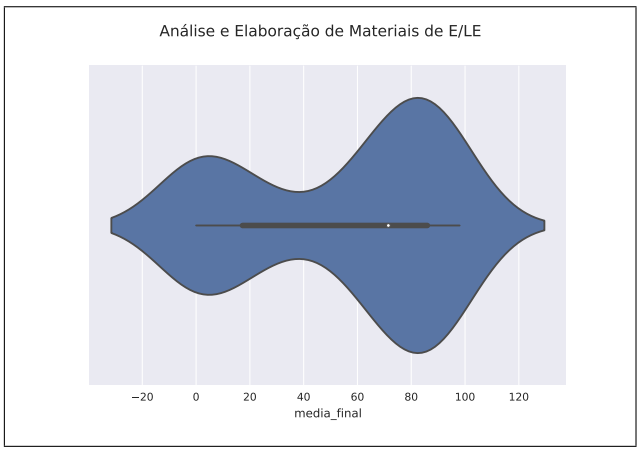
Entretanto, quando se trata de distribuições multimodais, o gráfico *boxplot* não consegue representar adequadamente a natureza do dado, uma vez que oculta a sua forma da distribuição, como pode ser observado na Figura 2.1, que representa os mesmos dados. A linha de densidade complementa as informações fornecidas pelo *boxplot* a partir da visualização gráfica da forma da distribuição dos dados. Para isso, ela é colocada simetricamente para esquerda e direita sobre o gráfico *boxplot*, dando origem ao gráfico tipo violino.

Figura 2.1: Comparação entre gráfico boxplot e violino.

(a) Exemplo de Gráfico Boxplot.



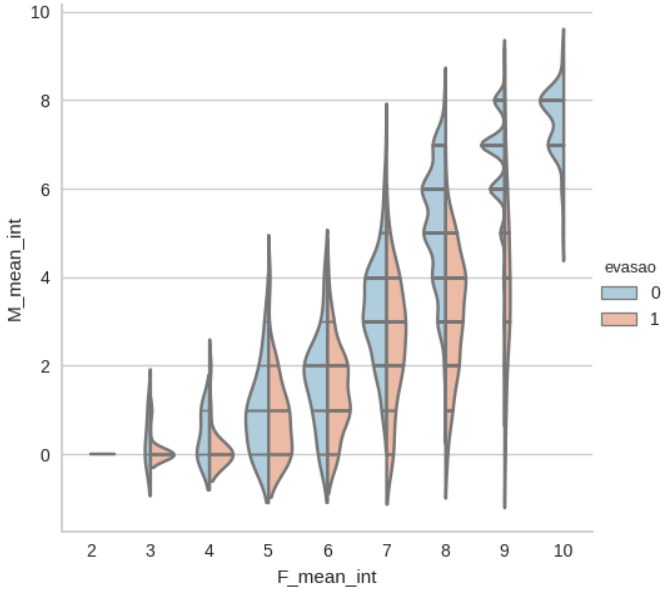
(b) Exemplo de Gráfico Tipo Violino.



Fonte: Elaborada pelo autor.

Os gráficos de violinos também podem ser utilizados para capturar a relação entre atributos, como visto na Figura 2.2, em que cada lado representa a distribuição de uma das variáveis.

Figura 2.2: Exemplo de gráfico tipo violino para relação entre atributos.



Fonte: Elaborada pelo autor.

2.2.2 Gráfico Q-Q

Se x e y são variáveis distribuídas de forma idêntica, então o gráfico de x-quartis versus y-quartis será uma linha reta com inclinação de 45 graus, apontada para a origem. Uma propriedade elementar dos gráficos de Q-Q é que se y for uma função linear de x, então o gráfico Q-Q correspondente ainda será linear, mas com possível mudança de localização e inclinação. Essa propriedade de invariância linear transforma o uso do gráfico Q-Q em uma abordagem valiosa, já que a linearidade é uma configuração geométrica a qual os seres humanos são capazes de perceber com mais facilidade (WILK & GNANADESIKAN 1968).

Para o caso em que as distribuições das variáveis têm caudas longas, o gráfico Q-Q tende a enfatizar a estrutura comparativa nas caudas, não dando tanta importâncias às distinções na parte central do gráfico, local onde as densidades são altas. A razão para isso é que o quartil é uma função de mudança rápida de probabilidade p quando a sua densidade é esparsa (nas caudas). Por outro lado, o quartil é uma função de mudança lenta de probabilidade p quando a sua densidade é alta (no meio).

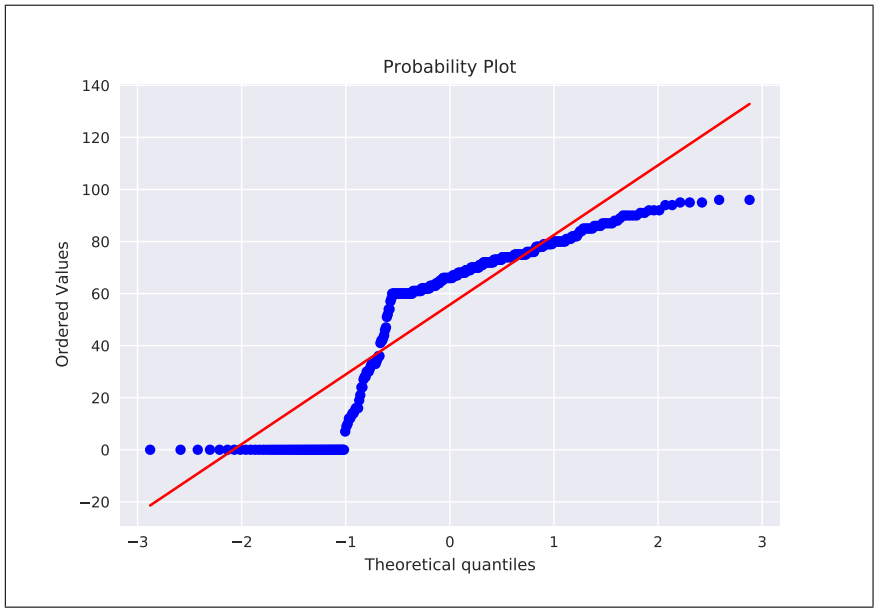
O gráfico Q-Q fornece um indicador muito sensível de discrepâncias entre duas distribuições e permite uma base útil para examinar a adequação de hipóteses compostas, em que os parâmetros não são especificados.

Como o gráfico Q-Q permite comparação entre as distribuições de dois conjuntos de dados observados, ele também é muito utilizado para comparações entre uma distribuição observada e a sua distribuição teórica (hipótese).

Uma característica interessante dos gráficos Q-Q é a verificação se uma distribuição

de dados segue uma distribuição normal e, em caso negativo, o possível motivo de sua discrepância. Quando o desvio da normalidade é provocado pelo desvio de assimetria, é observado arcos no gráfico Q-Q. Já quando o desvio é devido à mistura de distribuições, surgem fugas nos extremos do gráfico. Na Figura 2.3, é mostrado um exemplo de gráfico Q-Q que exibe a discrepância entre duas distribuições. Nesse gráfico Q-Q, pode-se observar que quanto mais distante de uma reta (em vermelho), maior é a discrepância entre as duas distribuições analisadas (em azul).

Figura 2.3: Gráfico Q-Q com fugas no extremo e arco.



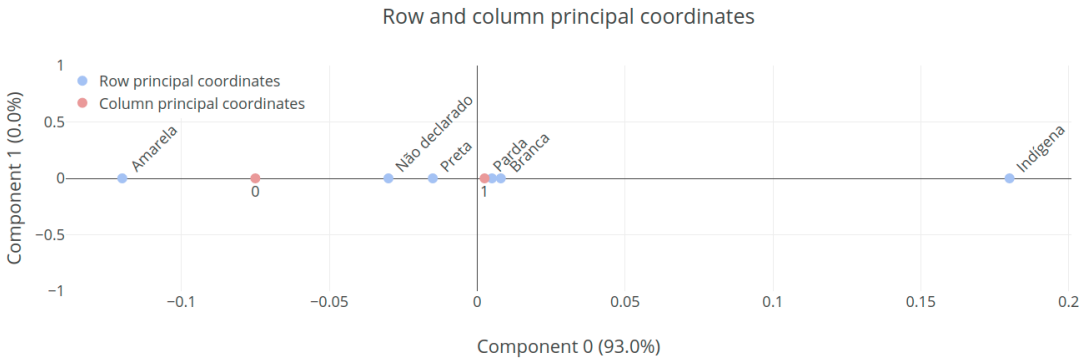
Fonte: Elaborada pelo autor.

2.2.3 Gráficos de análise de correspondência

A análise de correspondência (AC) é uma técnica exploratória em dados multivariados, frequentemente utilizada para redução de dimensionalidade e mapeamento perceptual em base composta por atributos categóricos (HAIR et al. 2009). O objetivo é esclarecer a relação entre os valores de duas variáveis categóricas dispostas em uma tabela de contingência a fim de descobrir uma explicação de baixa dimensão para possíveis desvios da independência dessas variáveis (Izenman 2008). A análise de correspondência se destaca pela construção do mapa perceptual a partir da associação de objetos descritos pelos atributos selecionados. Um exemplo de mapa perceptual é mostrado na Figura 2.4. Sua aplicação principal é exibir a correspondência entre categorias em escalas nominais e permitir representar duas variáveis categóricas em um mesmo diagrama. É importante destacar que essa técnica também pode ser utilizada em variáveis com valores contínuos, desde que sejam delimitados os limites, os valores das variáveis discretizadas e transfor-

mados em valores do tipo binário (cada valor possível da variável se torna uma coluna com valores 0 ou 1).

Figura 2.4: Exemplo de mapa perceptual.



Fonte: Elaborada pelo autor.

Para o cálculo da análise de correspondência, é inicialmente criada a tabela de contingência entre duas variáveis de escala nominal. A tabela de contingência representa a frequência de ocorrência conjunta entre os valores nominais de cada uma das duas variáveis. Após a criação da tabela de contingência, é realizado o cálculo do teste estatístico do qui-quadrado, a fim de padronizar os valores e gerar um índice de associação ou similaridade utilizado para criação do diagrama (mapa perceptual). O procedimento do cálculo do teste do qui-quadrado é descrito a seguir (HAIR et al. 2009):

1. **Valor esperado** - Esse valor representa o valor esperado para uma célula da tabela de contingência. O cálculo é feito a partir da probabilidade conjunta da combinação da coluna com a linha dessa tabela, através da probabilidade marginal para a coluna (total da coluna / total geral) vezes a probabilidade marginal para a linha (total da linha / total geral). Esse cálculo é descrito na equação a seguir:

$$FrequenciaEsperada = \frac{TotalColuna * TotalLinha}{TotalGeral} \tag{2.1}$$

2. **Diferença entre as frequências esperadas e reais** - Esse valor representa o quão distante o valor real está da frequência esperada naquela célula da tabela de contingência. A diferença é calculada a partir do passo anterior e os valores reais observados na tabela de contingência, sendo obtida via a equação a seguir:

$$Diferenca = FrequenciaEsperada - FrequenciaObservada \tag{2.2}$$

3. **Valor do teste qui-quadrado** - Esse valor é relacionado à intensidade de associação entre os valores nominais das variáveis. O cálculo é realizado a partir da razão entre a diferença ao quadrado calculada no passo anterior e a frequência esperada, que é

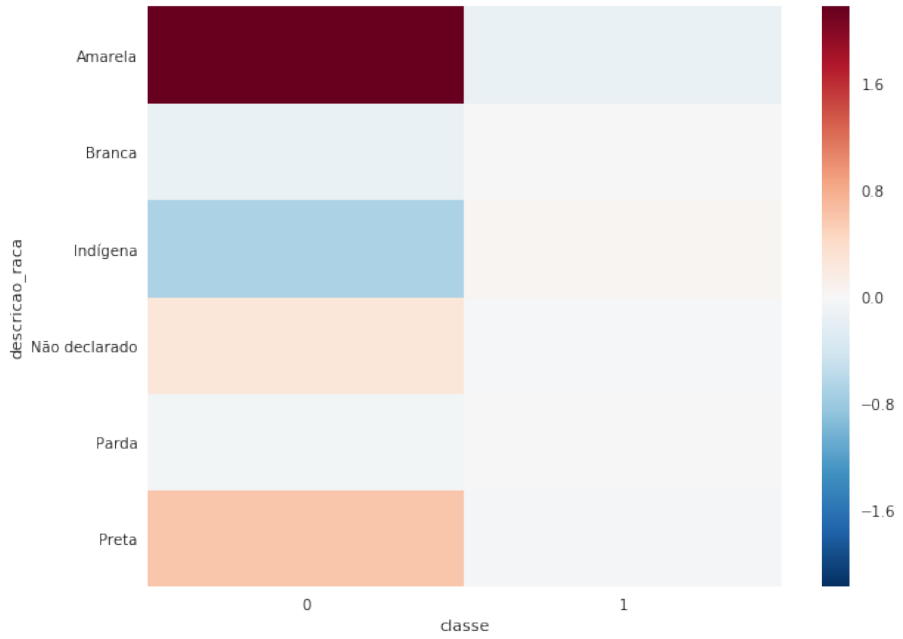
calculada no primeiro passo e descrita pela equação a seguir:

$$Qui - quadrado = \frac{Diferenca^2}{FrequenciaEsperada} \tag{2.3}$$

4. **Sinal da medida de similaridade**: O valor definido no teste qui-quadrado deve retomar a direção removida ao elevar a diferença ao quadrado no passo 3. Para tornar a medida mais intuitiva, o sinal inverso ao gerado deve ser atribuído no cálculo de diferença no passo 2. Portanto, caso o sinal tenha valor negativo, ele representa a repulsão entre os valores nominais de cada atributo. Por outro lado, caso o sinal seja positivo, o valor representa a força de atração entre os valores nominais de cada atributo.

Nas Figuras 2.4 e 2.5, são apresentados os valores gerados após o cálculo descrito acima entre dois atributos categóricos com valores nominais.

Figura 2.5: Exemplo de mapa de calor com os valores do teste do qui-quadrado.



Fonte: Elaborada pelo autor.

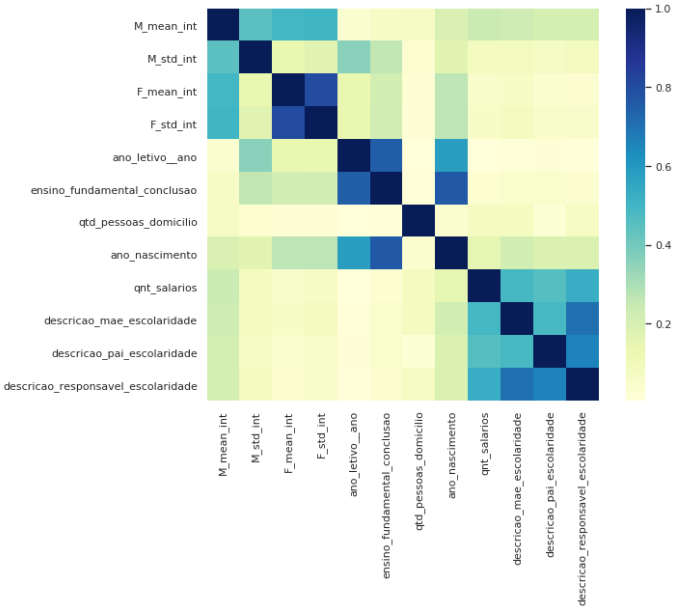
2.2.4 Matriz de correlação

A matriz de correlação (Ting 2017) é uma das técnicas mais utilizadas para visualização de correlações entre dados numéricos multivariáveis, pois permite identificar visualmente variáveis que se relacionam entre si. Essa matriz corresponde à normalização da Matriz de Covariância. É importante destacar que a determinação do grau de correlação entre duas variáveis depende da métrica adotada, sendo as mais utilizadas:

- **Correlação de Pearson** - Infere a relação linear entre as variáveis.
- **Correlação de Kendall** - É um teste estatístico não-paramétrico que infere a dependência entre as variáveis.
- **Correlação de Spearman** - Corresponde a um teste estatístico não-paramétrico que infere a associação entre as variáveis.

Os valores dos coeficientes gerados pelas métricas de correlações variam entre -1 e 1. O valor de +/- 1 indica a perfeita associação entre duas variáveis, sendo o $+$ uma relação positiva e o $-$ uma relação negativa. Já o valor 0 representa uma completa dissociação entre as duas variáveis. Na Figura 2.6, é apresentado um exemplo de matriz de correlação.

Figura 2.6: Exemplo de matriz de correlação entre atributos numéricos.



Fonte: Elaborada pelo autor.

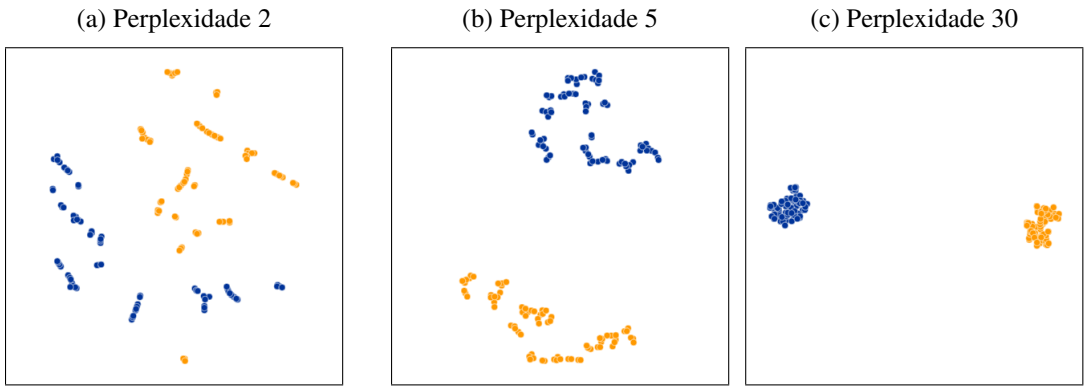
2.2.5 T-SNE: T-Distributed Stochastic Neighborhood Embedding

T-SNE é uma técnica comumente utilizada para redução de dimensionalidade de dados, principalmente para visualização de mapas bidimensionais (van der Maaten & Hinton 2008). A sua abordagem é não-linear e se adapta aos dados, executando diferentes transformações em diferentes regiões e preservando as estruturas locais dos dados enquanto converte dimensões mais altas para dimensões mais baixas. Desse modo, dados que estão próximos na alta dimensão também estarão próximos em baixa dimensão.

Além da quantidade de dimensões que deve ser definida na redução da dimensionalidade, outro importante parâmetro nessa abordagem é o parâmetro de perplexidade. Esse parâmetro indica o equilíbrio entre os aspectos locais e globais dos dados, além da quantidade de vizinhos próximos de cada ponto. É indicado que vários valores sejam testados,

sendo que os mais utilizados são valores entre 5 a 50 (Wattenberg et al. 2016). Na Figura 2.7, é apresentado o exemplo dos dados reduzidos para duas dimensões com diversos valores do parâmetro de perplexidade.

Figura 2.7: Exemplo de T-SNE com diferentes perplexidades.



Fonte: Adaptada de Wattenberg et al. (2016).

Na Tabela 2.1, comparamos os gráficos apresentados e a sugestão de como podem ser utilizados.

Tabela 2.1: Comparação entre técnicas de visualização de gráficos.

Nome	**Descrição**
Violino	Combinação do *boxplot* e linha de densidade. Interessante para visualização de distribuições multimodais e para comparar a distribuição entre atributos
Q-Q	Comparação de quartis entre duas distribuições. Interessante para verificar, de forma visual, normalidade, assimetria e se há mistura de distribuições
Análise de Correspondência	Os valores do teste Chi-Quadrado podem ser visualizados por diversas forma, como mapa de calor ou mapa perceptual. Interessante para realizar análise de independência entre atributos categóricos.
Matriz de Correlação	Matriz quadrada em que apresenta, de forma visual, a magnitude de uma correlação entre atributos par-a-par. Interessante para ser utilizado entre atributos numéricos.
T-SNE	Técnica para visualizar dados de forma bidimensional. Permite explorar agrupamentos nos dados de forma visual pelo humano. Também pode ser utilizado para redução de dimensionalidade.

Fonte: Elaborada pelo autor.

2.3 Pré-processamento dos dados

De posse da compreensão inicial levantada pela análise exploratória dos dados, é necessário realizar um conjunto de atividades para criação de uma base de dados mais adequada a ser utilizada pelos algoritmos de aprendizagem de máquina. A esse conjunto de atividades, chamamos de pré-processamento. As principais atividades nesse passo são a engenharia de atributos (composta por criação de novos atributos, preenchimento de dados ausentes, transformação dos dados), seleção de atributos e balanceamento dos dados. Nas seções a seguir, serão apresentados os principais algoritmos para cada uma dessas atividades.

2.3.1 Engenharia de atributos

A engenharia de atributos, *feature engeneering* em inglês (Domingos 2012), corresponde a um conjunto de técnicas com correspondentes algoritmos que visa tornar as informações contida nos dados de forma latente mais explícitas possíveis. Dentre essas técnicas, podemos citar:

- Criação de novos atributos;
- Preenchimento de dados ausentes; e
- Transformações de dados (discretização de dados contínuos e conversão em valores binários dos dados categóricos).

Para criação de novos atributos, o auxílio do especialista da área do problema a ser resolvido é fundamental para subsidiar a compreensão dos dados e validar a escolha dos atributos a serem criados. Nessa atividade, a criatividade é uma habilidade poderosa para o cientista de dados. Portanto, técnicas baseadas no *Design Thinking* (Filatro 2019) podem ser bem aplicadas para uma melhor sistematização do processo de criação de atributos. Além do esforço criativo de novos atributos, algumas ferramentas emergentes prometem facilitar a criação automática de novos atributos a partir dos dados existentes (Kanter & Veeramachaneni 2015).

Preenchimento de dados ausentes

O problema de dados ausentes (faltantes) é algo comum em diversos contextos de sistemas baseados em dados. Por exemplo, nas áreas de educação e psicologia, há uma taxa de falta de dados entre 15% a 20% (Enders 2003). Quando esse problema ocorre, três grandes consequências surgem: as instâncias com falta de valores são sistematicamente diferentes das unidades com dados completos, consequentemente análises ingênuas que ignorar essas diferenças podem ser tendenciosas; a existência de dados ausentes geralmente implica em perda significativa de informações, logo as estimativas podem ser menos eficientes do que o planejado inicialmente; e métodos estatísticos padrões são projetados para conjuntos de dados completos (Little & Schenker 1995).

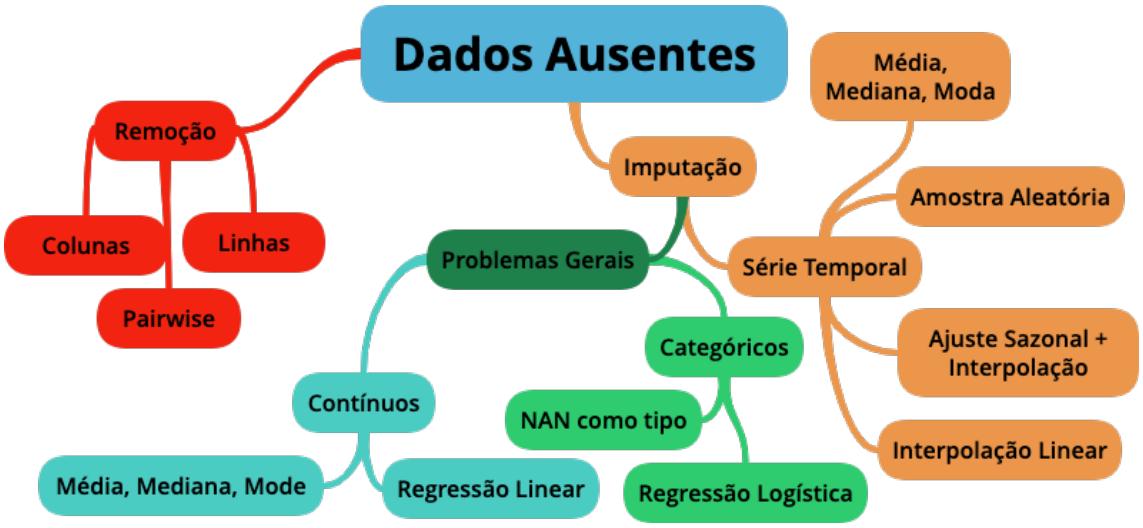
Uma das soluções para o problema de dados ausente nas bases de dados é o seu preenchimento. Técnicas que utilizam dessa abordagem são conhecidas como técnicas de

imputação de dado (*data imputation*, em inglês). Ao utilizar essa abordagem, é importante que os dados gerados mantenham a distribuição estatística original, para não comprometer o desempenho dos resultados. Além disto, esse preenchimento não é simples, pois pode haver a valorização de vieses, preconceitos e ruídos que não são condizentes com a realidade, principalmente se houver poucas instâncias na base de dados (Kahneman & de Arantes Leite 2012, Raschka 2018). Por exemplo, utilizar a média para preencher a renda familiar nos campos faltantes de uma base de dados pode tornar os dados mais enviesados para uma renda menor do que o real uma vez que, em geral, há um quantitativo maior de famílias que recebem de um a dois salários mínimos, puxando a média para baixo. Continuando o exemplo anterior, outra estratégia mais sofisticada é utilizar de conhecimentos sobre o contexto, como por exemplo o fato de que pessoas com maior escolaridade tendem a ter maior renda (Salvato et al. 2010). Nesse caso, é interessante utilizar a média de acordo com a escolaridade para a imputação de dados faltantes.

Outra forma de trabalhar com os dados ausentes é a remoção do atributo quando há um quantitativo muito alto de dados faltante, a ponto de comprometer a qualidade desse atributo. Não há uma porcentagem definida de dados ausentes para exclusão do atributo, inclusive há pesquisas que argumentam para não utilizar essa porcentagem como guia (Madley-Dowd et al. 2019). Entretanto, essa decisão geralmente é feita a partir do contexto, da experiência do cientista de dados e dos testes empíricos.

Na Figura 2.8, diversas técnicas de como tratar os dados ausentes em base de dados segundo (Swalin 2018) são sistematizadas.

Figura 2.8: Sistematização para tratar dados ausentes em bases de dados.



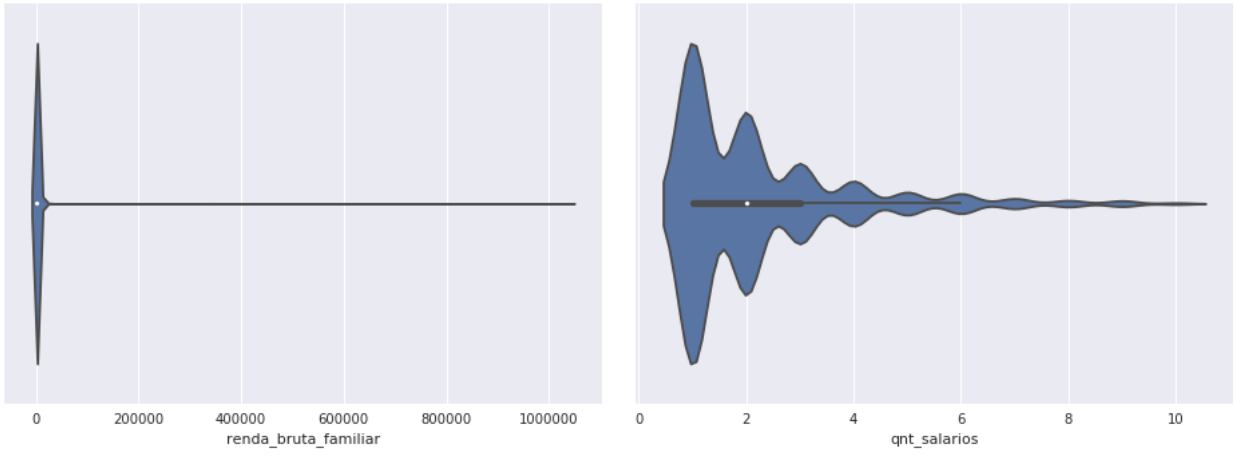
Fonte: Adaptada de Swalin (2018).

Transformação dos dados

Quando formulamos modelos baseados em dados, geralmente assumimos algumas premissas, como linearidade, ausência de ruído, ausência de colinearidade, atributos com distribuição normal e atributos em escalas com ordem de grandezas similares. Entretanto, em sistemas reais, os dados disponíveis geralmente não estão nessas condições. Portanto, são necessárias algumas transformações para que os dados estejam mais adequados para a geração de bons modelos.

Além de transformações para tornar os dados mais adequados aos modelos, há outras modificações referentes à experiência do cientista de dados e do contexto. Por exemplo, no domínio educacional, o humor do professor pode influenciar nas notas dos alunos (Brackett et al. 2013). Logo, para tentar minimizar esse problema, em vez de utilizar o valor da nota com duas casas decimais, podem ser utilizado valores discretos como os conceitos ou o valor inteiro de 1 a 10. Outro exemplo de discretização se aplica aos valores de renda. Para o contexto educacional do IFRN, por exemplo, sabe-se que a grande maioria dos alunos possui renda familiar de um a dois salários mínimos e uma quantidade minoritária de alunos possui renda familiar superior a dez salários mínimos. Logo, podem ser criado valores discretos de 1 a 9 salários mínimos e todos acima de 9 serão inclusos no grupo com renda familiar acima de 10 salários mínimos, como visto na Figura 2.9.

Figura 2.9: Exemplo de discretização de dados. No lado esquerdo, valores de renda bruta de 0 a 1 milhão. No lado direito, os valores discretizados em até 10 salários mínimos.



Fonte: Elaborada pelo autor.

Outra importante transformação sobre os dados diz respeito aos do tipo categórico. A maioria dos modelos de aprendizagem trabalha com valores numéricos. Logo, é necessário o emprego de técnicas que transformem os dados do formato categórico para o formato numérico, mas mantendo características importantes, como a falta de ordem entre os valores categóricos nominais. Como exemplo, podemos citar o fato que as cores não possuem relação de maior ou menor entre elas. Para isso, umas das técnicas mais

utilizadas é a transformação de dados categóricos para dados do tipo binário. Essa técnica de codificação cria um atributo fictício para cada valor distinto que o valor categórico pode assumir. Depois que os valores fictícios são criados, um valor booleano (0 ou 1) é preenchido para indicar se o valor é verdadeiro ou falso para uma dada instância do atributo. Como consequência, uma matriz ampla e esparsa é obtida, cujos elementos podem assumir valores 0 ou 1.

Na Figura 2.10, é apresentado um exemplo de transformação de uma variável categórica nominal originalmente com 4 valores.

Figura 2.10: Exemplo de dado na forma binária do atributo conceito.

	conceito_S	conceito_I	conceito_R	conceito_O
0	0	0	0	0
1	0	0	1	0
10	0	0	0	0
1000	1	0	0	0
1001	0	0	1	0

Fonte: Elaborada pelo autor.

2.3.2 Seleção de atributos

Como visto na seção anterior, a criação de novos atributos é uma das etapas mais importantes para extrair informações latentes dos dados. Entretanto, o aumento da dimensionalidade e o esparsamento da base de dados são o efeito colateral dessa etapa, principalmente quando ela contém um grande quantitativo de atributos categóricos, podendo ocasionar o problema conhecido como "Maldição da Dimensionalidade"(Verleysen & François 2005). Diante disso, técnicas para redução de dimensionalidade dos dados podem evitar esse problema e gerar modelos de aprendizagem mais simples de serem entendidos pelos humanos. Além disso, a redução do número de atributos pode levar a vários benefícios (Ippolito 2019), como:

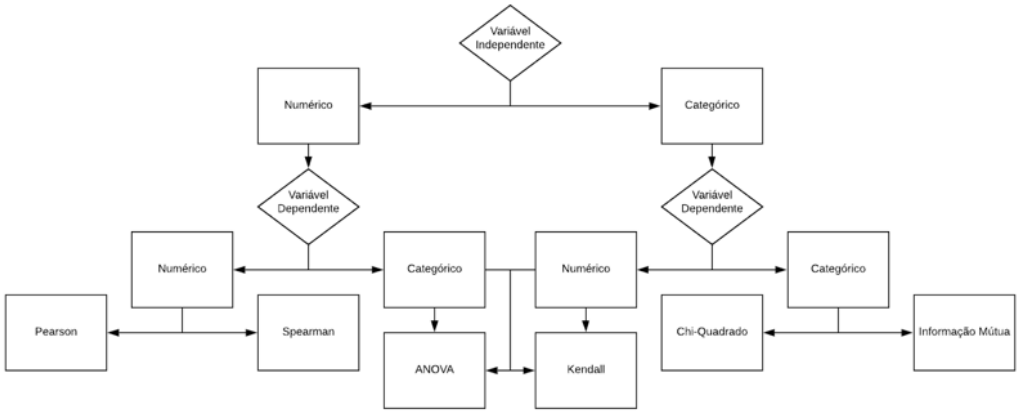
- Melhorias de precisão;
- Redução de risco de *overfitting*;
- Aceleração no treinamento;
- Visualização de dados aprimorada e
- Aumento da interpretação do modelo.

Há diferentes formas de se categorizar os algoritmos de seleção de atributos, sendo a por perspectiva uma das mais utilizadas, em que são criados três grupos de algoritmos: por filtro (em inglês, *Filter Method*), por empacotamento (em inglês, *Wrapper Method*) e embutido (em inglês, *Embedded Method*) (Li et al. 2017).

O método por filtro encontra o conjunto de atributos mais relevantes de acordo com alguma métrica independente do modelo, mas relacionada com o tipo (numérico ou categórico) de variável dependente e independente (Brownle 2019), como sistematizado na Figura 2.11. Para um problema de classificação binária, destacamos os testes estatísticos ANOVA e o Chi-Quadrado (descrito na Seção 2.2.3). O primeiro é recomendado quando as variáveis independentes são numéricas, já o segundo quando essas variáveis são categóricas. O ANOVA é um teste estatístico usado para verificar as médias de dois ou mais grupos que são significativamente diferentes uns dos outros. Para a seleção de atributos, realiza-se o teste estatístico F, que assume como hipótese nula (H0) para cada atributo numérico independente que o mesmo tem variância igual ao da classe (variável dependente). Caso a hipótese nula seja rejeitada, significa que o atributo tem impacto sobre a classe; logo, o atributo é selecionado.

Esses algoritmos são os mais eficientes computacionalmente, mas não garantem o conjunto ótimo para um dado modelo. Outra aplicação para os métodos por filtros são remover atributos que ocasionam problemas aos modelos. Por exemplo, a multicolinearidade é um caso de regressão múltipla em que os atributos são altamente correlacionados. A partir do uso de técnicas como a variância do fator de influência (VIF) (Paul 2008), pode-se detectar a presença de multicolinearidade, levando à sua devida remoção (Katrutsa & Strijov 2017).

Figura 2.11: Métricas estatísticas para seleção de atributos pelo método filtro

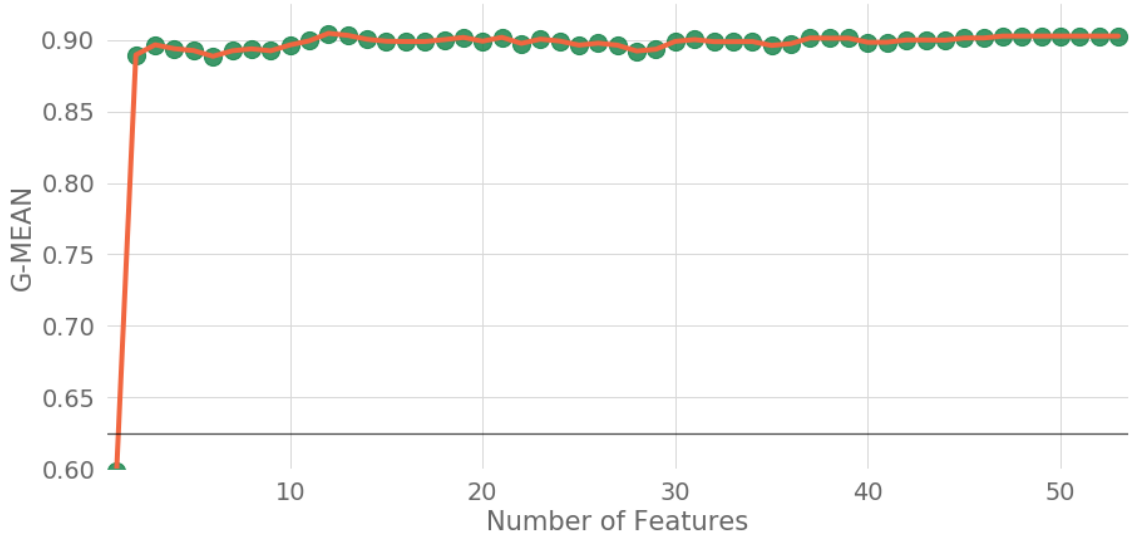


Fonte: Adaptado de Brownle (2019)

Os métodos de *Wrapper Method* encontram o conjunto de atributos mais relevantes a partir do critério de avaliação do modelo. Nele, podem ser realizadas a remoção (*Backward*) ou inclusão (*Forward*) do atributo de acordo com a avaliação do seu impacto sobre a métrica de avaliação do modelo. Um dos algoritmos mais conhecidos desse tipo de método de seleção de atributos é o *Recursive Feature Elimination and Cross-Validated Selection* (RFEC) (Guyon et al. 2002), que seleciona os melhores atributos a partir da eliminação recursiva deles, para uma dada métrica de avaliação. Mais especificamente,

a cada iteração do algoritmo, é eliminado um atributo e verificado se há diminuição do desempenho do modelo preditivo (Figura 2.12). Se o desempenho do modelo diminuir significa que o atributo era importante, então ele retorna ao conjunto de dados; caso contrário, o atributo é excluído do conjunto de dados. Esse algoritmo necessita definir, *a priori*, qual o modelo baseado em dados que será utilizado para avaliação.

Figura 2.12: Exemplo do RFEC



Fonte: Elaborada pelo autor

Por fim, os métodos do tipo *Embedded Method* selecionam os atributos a partir de um modelo de forma similar aos *Wrapper*, porém mais eficiente, pois não precisa avaliar os atributos iterativamente como, por exemplo, no RFEC. A técnica mais conhecida desse tipo de método de seleção de atributos é a baseada em modelos de regularização (LASSO), que trata a seleção de atributos como se fosse um problema de otimização com restrições, com o objetivo de minimizar os erros de aprendizagem com o modelo menos complexo possível.

2.3.3 Técnicas de balanceamento de dados

Um dos problemas que afeta fortemente o desempenho de modelos de classificação de dados é decorrente de classes desbalanceadas. Nesse cenário, as quantidades de amostras disponíveis correspondentes a cada classe são muito desproporcionais entre si. Para o caso de um modelo de classificação binária, teríamos uma classe majoritária e outra minoritária. Nesse contexto, pode ocorrer o fenômeno conhecido como "Paradoxo da Acurácia"[1], que é uma situação em que um alto valor de acurácia não corresponde a um modelo de alta qualidade (He & Ma 2013, Zhu & Davidson 2007). Por exemplo, se um determinado

[1] Acurácia corresponde à taxa de acerto do modelo em classificar corretamente os dados

conjunto de dados incluir 1% dos exemplos de classe minoritária e 99% dos exemplos de classe majoritária, uma abordagem ingênua para classificar cada exemplo como classe majoritária fornecerá uma acurácia de 99%. No entanto, esse resultado não reflete o fato de que nenhum dos exemplos minoritários foi identificado. É essencial destacar que, em muitas situações, esses exemplos minoritários são mais importantes que a classe majoritária, como os problemas relacionados a doenças graves (câncer, por exemplo) e problemas educacionais (como evasão escolar).

Na literatura, há basicamente duas abordagens convencionais para se resolver o problema de dados desbalanceados no contexto de classificação: (I) aprendizado sensível a custos (*cost-sensitive*, em inglês), que atribui um alto custo à classificação incorreta da classe minoritária e tenta minimizar o custo total; e (II) a técnica de amostragem, que consiste em criar um conjunto de dados com uma distribuição de classe apropriada (Chen et al. 2004).

As duas estratégias mais populares de técnicas de amostragem são a *downsample* (também conhecida como *undersample*) e a *upsample* (Chen et al. 2004). Na *downsample*, as instâncias da classe majoritária são descartadas aleatoriamente até que uma distribuição dos dados mais equilibrada seja atingida (Figura 2.13). Considere, por exemplo, um conjunto de dados com dez instâncias da classe minoritária e 90 instâncias da classe majoritária. Na estratégia *downsample*, será criado um balanceamento entre as distribuição das duas classes a partir da remoção de 80 instâncias da classe majoritária. Assim, o conjunto de dados resultante consistirá em 20 instâncias, com dez instâncias da classe majoritária (selecionadas aleatoriamente) e dez instâncias da classe minoritária iguais ao conjunto de dados original.

No caso da estratégia *upsample*, instâncias das classes minoritárias são copiadas e repetidas no conjunto de dados original até que seja alcançada uma distribuição mais equilibrada entre as classes (Figura 2.13). Nesse caso, se houver duas instâncias da classe minoritária e 100 instâncias da classe majoritária, as duas instâncias da classe minoritária serão copiadas 49 vezes cada. O conjunto de dados resultante consistiria em 200 instâncias: a classe majoritária com 100 instâncias e 100 instâncias de classe minoritária (50 cópias de cada uma das duas instâncias de classe minoritária).

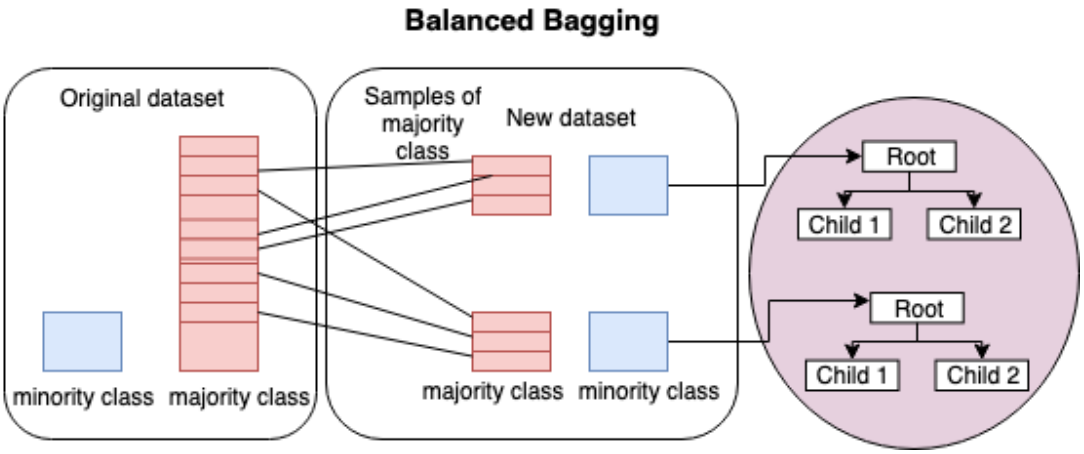
Figura 2.13: Fluxo para balanceamento de dados a partir da técnica de amostragem.

Fonte: Elaborada pelo autor.

Embora as estratégias de *downsample* e *upsample* produzam bons resultados, ambas possuem deficiências. Por exemplo, na *downsample* várias instâncias são descartadas; logo, instâncias que poderiam produzir uma melhor superfície de decisão podem ser perdidas no processo de descarte de amostras. Em relação à *upsample*, copiar as instâncias minoritárias pode causar *overfitting*, fazendo com que o desempenho do classificador se deteriore (He & Ma 2013).

Para evitar o problema de *overfitting* no uso de estratégia *upsample*, duas técnicas podem ser usadas: SMOTE e ADASYN. A técnica SMOTE (*Synthetic Minority Oversampling Technique*, em inglês) é a versão mais usada e seu algoritmo consiste em copiar dados usando a estratégia baseada no algoritmo *K-Nearest Neighbours* (KNN) (He & Garcia 2009). Já a ADASYN utiliza o conceito de usar uma distribuição ponderada para diferentes exemplos da classe minoritária, em que mais dados sintéticos são gerados para os exemplos da classe minoritária mais difíceis de aprender (Haibo He et al. 2008). Também existem abordagens híbridas com excelentes resultados como a *Balanced Bagging*, cujo algoritmo inicialmente cria novos subconjuntos a partir de amostras do conjunto original, com quantidades igualitárias de instâncias entre a classe minoritária e a classe majoritária. Na próxima etapa, para cada um desses subconjuntos, o algoritmo treina uma árvore de decisão. Ao final, agrega as predições de cada uma das árvores (Chen et al. 2004, Lemaître et al. 2017), como visto na Figura 2.14.

Figura 2.14: Fluxo de balanceamento de dados realizado pelo próprio modelo de aprendizagem.



Fonte: Elaborada pelo autor.

Outras abordagens para mitigar o problema de dados desbalanceados são o uso de modelos que contemplem no treinamento o balanceamento como o *Balanced Cascade* (Liu et al. 2006); o SVM atribuindo custos às instâncias (Witten et al. 2016); a Naive-Bayes, uma vez que as previsões são calibradas pela probabilidade da amostra; as árvores de decisão geradas a partir do cálculo da distância de Hellinger (HDDTs) (He & Ma 2013); k-means *balancing* (Rieger et al. 2014) e amostragem probabilística (Grósz et al. 2017, Lawrence et al. 2012).

Na Tabela 2.2, estão sumarizadas as principais características das técnicas de balanceamento apresentadas.

Tabela 2.2: Comparação entre técnicas de Balanceamento.

Nome	**Descrição**
Downsample	Técnica simples baseada em amostragem com baixo custo computacional. Reduz a base de dados. Pode piorar a construção da superfície de decisão devido ao descarte de algumas instâncias.
Upsample	Técnica simples baseada em amostragem com baixo custo computacional. Aumenta a base de dados. Pode causar *overfitting* devido ao excesso de cópias. Para evitar esse problema, há algoritmos derivados como SMOTE e ADASYN
Balanced Bagging	Balanceamento realizado internamente pelo próprio algoritmo de aprendizagem. Computacionalmente custoso. Em regra, apresenta bons resultados. Mantém o tamanho da base.

Fonte: Elaborada pelo autor.

2.4 Modelos baseados em aprendizagem de máquina

Nas seções anteriores, foram apresentadas as várias técnicas e algoritmos de pré-processamento e análise exploratória de dados, que constituem atividades importantes num ciclo de geração de modelos baseados em dados. Nesta seção, será tratada a parte de construção do modelo propriamente dito. No nosso caso, iremos abordar especificamente modelos baseados em técnicas de aprendizagem de máquinas supervisionadas.

Os modelos baseados em técnicas de Aprendizagem de Máquina (AM) (*Machine Learning*, em inglês) são aqueles produzidos a partir de algoritmos de aprendizagem que tomam como base o histórico dos dados associados ao fenômeno sob estudo (Faceli et al. 2011). O termo "Aprendizagem de Máquina"foi cunhado em 1959 por Arthur Samuel, viabilizando um trabalho que "dá aos computadores a capacidade de aprender sem serem explicitamente programados"(Samuel 1959).

As técnicas de AM podem ser dividas em: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Na aprendizagem supervisionada, os dados de treinamento do modelo consistem no conjunto de atributos e um rótulo (ou valor alvo) associado a eles. O objetivo dos algoritmos dessa classe é produzir um modelo, hipótese, ou função capaz de relacionar os atributos de entrada com o atributo alvo. Por exemplo, podemos ter como atributos de entrada informações sobre alunos (como caracterização social, demográficos e educacionais) e como valor alvo o seu desempenho na disciplina de português. Ao entrar um novo grupo de alunos na instituição de educação, podemos tentar prever o desempenho deles na referida disciplina de acordo com o modelo gerado (treinado) previamente com os dados disponíveis anteriormente. Quando o atributo alvo é do tipo numérico, temos um processo de regressão e, caso o atributo alvo seja do tipo categórico, temos um processo de classificação de dados.

Já na aprendizagem não supervisionada, não há atributo alvo que deve ser inferido. Isso nos permite abordar problemas com pouca ou nenhuma ideia sobre os resultados. Algoritmos de aprendizagem não supervisionadas geralmente são usados para encontrar padrões e agrupamentos nos dados. Por fim, as técnicas de aprendizado por reforço consistem em tomar ações adequadas para maximizar a recompensa em uma situação específica. Geralmente essa abordagem de aprendizagem por reforço é utilizada para encontrar o melhor comportamento ou caminho possível a seguir em uma situação específica. Na Figura 2.15, é apresentado um infográfico dessas três abordagens, com suas respectivas aplicações (Chakure 2019).

Figura 2.15: Infográfico sobre abordagens e aplicações de aprendizagem de máquina.

Fonte: Adaptado de Chakure (2019)

Ressaltamos que, nesta tese, utilizaremos técnicas supervisionadas de aprendizagem de máquina para a geração dos nossos modelos. Mais especificamente, iremos gerar modelos de classificação binária, em que as características socioeconômicas e de desempenho escolar dos alunos são os atributos de entrada (variáveis independentes) e o status da matrícula dos alunos (persistente ou evadido) é o atributo alvo no modelo (variável dependente).

Há diversos algoritmos de aprendizagem de máquina supervisionados inspirados em diferentes área de conhecimento e heurísticas. Vários desses algoritmos estão implementados em pacotes em diversas linguagens de programação, como *Scikit-learn* (Pedregosa et al. 2011). A seguir, são descritos brevemente o funcionamento conceitual de alguns desses algoritmos (Igual & Seguí 2017), (Cady 2017), (Faceli et al. 2011), (Witten et al. 2016) (HAYKIN 2001).

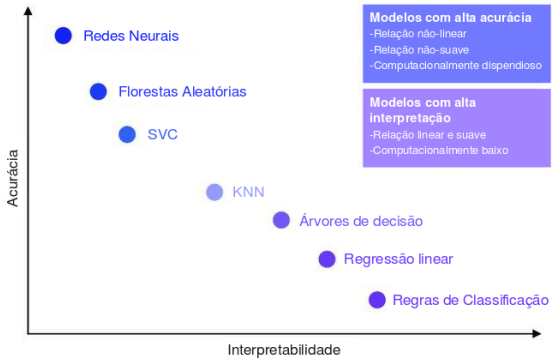
- **Árvores de Decisão**: Árvores de decisão (em inglês, *Decision Tree* - DT) é um método de aprendizagem supervisionada não paramétrico usado para classificação e regressão. O objetivo é criar um modelo baseado em regras de decisão simples inferidas a partir da estratégia de dividir para conquistar. O objetivo da árvore de decisão é dividir os dados com base na sua homogeneidade. Geralmente, para avaliar a homogeneidade de uma divisão, é realizado o cálculo da Entropia (também conhecido como ganho de informação) ou o índice de Gini (também conhecido como impureza). A principal vantagem das DT é que elas geram modelos simples de fácil interpretação, porém sua maior deficiência reside no fato de serem bastante sensíveis aos dados, podendo apresentar um erro alto de variância (vide seção 2.5).
- **Regressão Logística**: a Regressão logística (em inglês, *Logistic Regression* - LR) é o modelo que calcula a probabilidade de uma instância pertencer a uma determinada classe. Para esse modelo, é importante que não haja multicolinearidade entre os atributos e que as ordens de grandeza deles sejam similares. Esse modelo funciona bem quando há uma relação linear entre os atributos de entrada e o atributo alvo. A busca pelo hiperplano que separa as classes tem chance de tornar o modelo incapaz de capturar a complexidade de padrões sofisticado, podendo gerar um alto erro de *bias* (vide Seção 2.5). Entretanto, esse mesmo motivo ajuda a evitar a captura de padrões aleatórios (evitando o erro de variância).
- ***Naive Bayes***: os métodos *Naive Bayes* são um conjunto de algoritmos de aprendizado supervisionados baseados na aplicação do teorema de Bayes com a suposição "ingênua" de independência entre os atributos de entrada do modelo.
- **KNN**: KNN é o acrônimo de *K-Nearest Neighbors* em inglês, ou k-vizinho mais próximo, em português. O princípio por trás dos métodos de KNN é encontrar um número pré-definido de amostras de treinamento mais próximas do novo ponto e prever o rótulo a partir delas. O número de amostras pode ser uma constante definida pelo usuário (k) ou variar com base na densidade local de pontos (aprendizado de vizinho baseado em raio). A distância pode, em geral, ser qualquer medida métrica, sendo a distância euclidiana a escolha mais comum.
- **Redes neurais artificiais**: Uma rede neural artificial é um processador massivamente distribuído paralelo, composto de unidades simples de processamento, que têm a propensão natural de armazenar conhecimento experimental e disponibilizá-lo para uso. As redes neurais (principalmente as *Deep Learning*) conseguem trabalhar com uma quantidade enorme de dados e extrair os padrões complexos. Entretanto, os modelos dessas redes podem cair no problema de erro alto de variância (vide seção 2.5).
- **Florestas Aleatórias**: conhecido como *Random Forest* (RF) em inglês, consiste um conjunto de vários classificadores de árvore de decisão treinado a partir de várias subamostras do conjunto de dados original. O modelo RF usa a média para melhorar a precisão preditiva e controlar o *overfitting* no processo de treinamento do modelo. O tamanho da subamostra é sempre o mesmo que o tamanho da amostra da entrada original e seus pontos são escolhidos a partir da seleção aleatória com substituição. As RF conseguem capturar padrões de dados complexos, entretanto podem cair no problema de um alto erro de variância (vide seção 2.5).

- **SVC**: as máquinas de vetores de suporte (em inglês, *Support Vector Machines* - SVM) são embasadas pela teoria de aprendizado estatístico, em que se assume a separação linear dos dados a partir de um hiperplano ótimo de separação.
- **SGDC**: Esse algoritmo implementa modelos lineares com restrição e o uso de gradiente estocástico (SGD, acrônima para *Stochastic Gradient Descent* em inglês). O gradiente da perda é estimado em cada amostra e o modelo é atualizado ao longo do caminho com taxa de aprendizado decrescente.

Esses algoritmos podem ser classificados como caixa-branca ou caixa-preta, de acordo com a capacidade de compreensão pelo humano do modelo gerado, sendo o de caixa-branca com maior compreensão e o de caixa-preta o oposto. Entretanto, existe uma relação inversa entre a acurácia do modelo (quanto mais complexo, melhor a acurácia) e sua facilidade de interpretação por humanos (quanto mais simples, melhor a interpretação). Conforme visto na Figura 2.16 baseada em (Morocho-Cayamcela et al. 2019) e (James et al. 2014), é observado que dois modelos com fácil interpretação são a Regressão e a Árvore de Decisão.

Isso ocorre devido à troca inerente entre a flexibilidade e a interpretabilidade de um modelo. Flexibilidade é a característica do modelo de ajustar seus parâmetros para mapear uma função e, dessa forma, quanto mais flexível for um modelo, melhor o ajuste pode ser produzido, o que aumenta sua precisão preditiva. Entretanto, um modelo mais flexível é geralmente mais complexo e requer mais parâmetros para o ajuste, tornando mais difícil a sua interpretação. Por outro lado, os parâmetros em um modelo linear são relativamente simples e interpretáveis, apesar de terem dificuldades em encontrar padrões complexos. Como se pode perceber, os modelos de aprendizado de máquina mais flexíveis e com melhor precisão tendem a ser menos interpretáveis pelos humanos (James et al. 2014).

Figura 2.16: Modelos de acordo com acurácia e interpretação.



Fonte: Adaptado de Morocho-Cayamcela (2019).

Para o contexto do problema de evasão escolar, os EWS geralmente funcionam como uma caixa-preta, simplesmente emitindo alertas de desempenho, mas não fornecendo meios para entender ou rastrear a motivação das predições. Consideramos esse aspecto como uma das principais deficiências dos EWS, pois sem conhecer os fatores de risco,

os educadores não podem adaptar ou melhorar os currículos, nem fornecer uma intervenção personalizada para o aluno (Cano & Leonard 2019). Então, em problemas nos quais se faz necessário a compreensão do fenômeno pelo humano para a adequada tomada de decisão, é preferível optar-se por técnicas que gerem modelos com maior capacidade de interpretação, como as árvores de decisão e regressão logística (Cady 2017). Técnicas sobre interpretabilidade de modelos de dados baseados serão discutidos na seção 2.6.

2.5 Avaliação de modelos de aprendizagem supervisionada

Uma vez estabelecida qual técnica de aprendizagem de máquina supervisionada será utilizada para a geração do modelo de classificação de dados, é necessário definir como eles serão avaliados e comparados. Há três aspectos da avaliação importantes que devem ser analisados:

1. Estimar o desempenho do modelo;
2. Seleção dos melhores hiper-parâmetros para um modelo;
3. Comparação entre os modelos treinados.

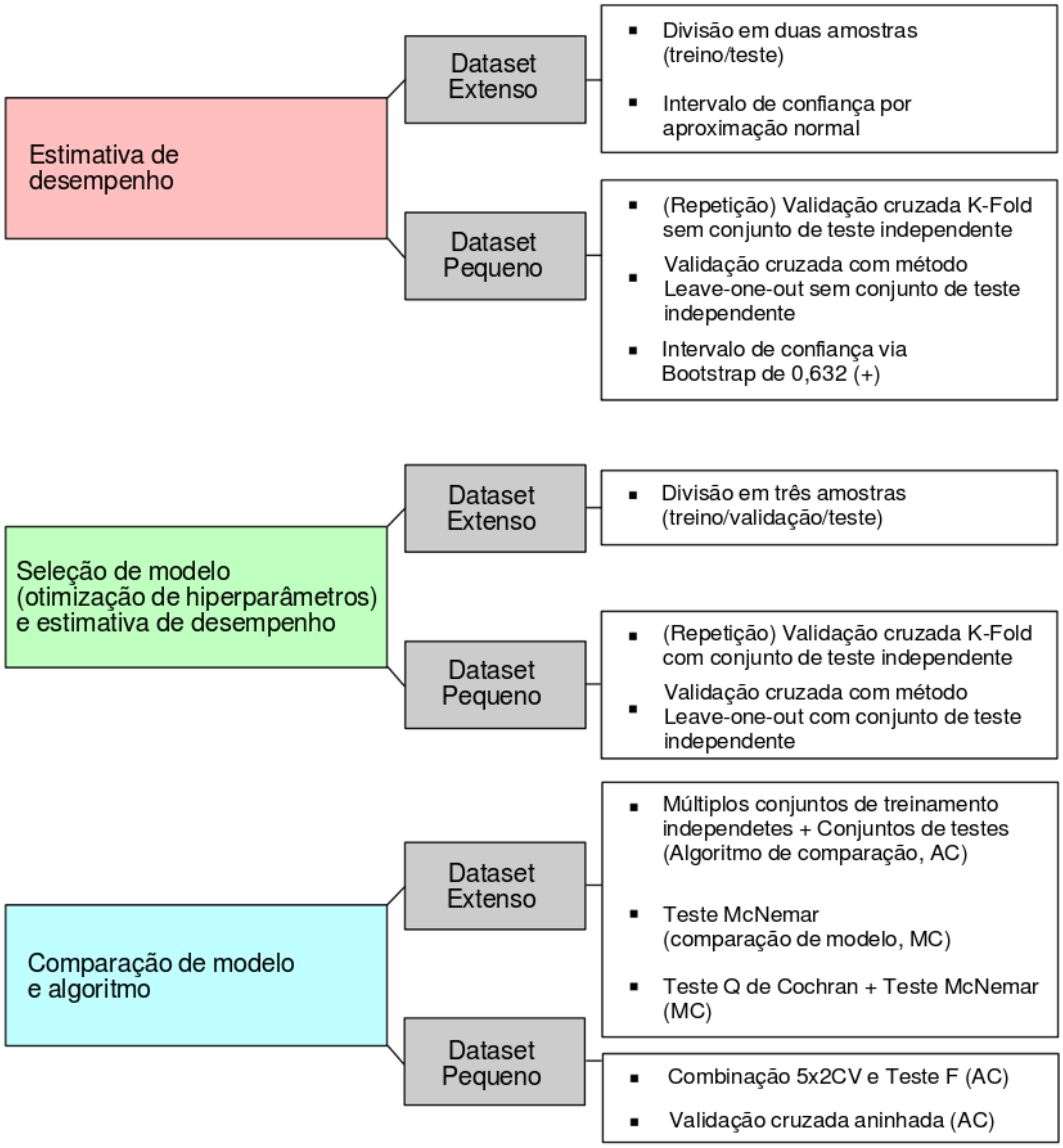
Estimar o desempenho do modelo diz respeito ao valor de uma dada métrica de avaliação (seção 2.5.2) de um modelo de aprendizagem sobre um conjunto de teste. A seleção dos melhores hiper-parâmetros está relacionada ao ato de identificar qual o conjunto de hiper-parâmetros de um dado modelo de aprendizagem tem o melhor desempenho para uma métrica definida sobre um conjunto de dados de treinamento. Essa identificação pode ser realizada por uma busca exaustiva da melhor combinação de parâmetros, ou utilizando alguma heurística. Por fim, comparar os modelos é o ato de verificar que dado dois modelos de aprendizagem, qual de fato é superior diante de com conjunto de testes para um dado teste estatístico. Diversas técnicas são utilizadas em cada uma dessas fases, como visto na Figura 2.17 (Raschka 2018).

Diante dessa classificação, nas próximas subseções, serão apresentadas algumas estratégias e métricas para estimar desempenho de modelos, além de testes estatísticos para avaliar desempenho comparativa entre modelos.

2.5.1 Estratégias para estimar desempenho do modelo

Para estimar o desempenho do modelo, é realizada uma estratégia de separação dos dados, dividindo em conjunto de treinamento e conjunto de teste. Ressalta-se que o processo de treinamento e validação do modelo é realizado de forma *offline*. O conjunto de treinamento é utilizado para o aprendizado do modelo e otimização dos parâmetros. O grupo de teste é um conjunto de instâncias que não foram apresentadas no treino, a fim de verificar o poder de generalização do modelo. Geralmente são as métricas utilizadas no conjunto de testes que são levadas em consideração na avaliação final do modelo. A proporção dessa divisão não é estabelecida, mas os valores mais usuais de treino e

Figura 2.17: Fluxograma com recomendações de qual técnica utilizar para comparar modelo (MC) ou algoritmo (AC)



Adaptado de Raschka (2018)

teste respectivamente são 60/40, 70/30, 80/20 ou até 90/10, se a base de dados for relativamente grande (Raschka 2018). A decisão geralmente perpassa pela quantidade de instâncias disponíveis, pela relação com a quantidade de atributos e pela experiência do cientista de dados.

Além de avaliar de fato os valores de métrica, é importante verificar as curvas de aprendizagem, a fim de analisar o erro de *bias* e variância (Domingos 2012):

1. Erro de *Bias*: também conhecido como *underfitting*, ocorre quando o modelo não captura os padrões nos dados, ou seja, o modelo não apresenta a complexidade necessária para aprender os padrões. Para solucioná-lo, modelos de aprendizagem mais complexos devem ser utilizados ou a coleta de mais dados sobre o problema deve ser feita.
2. Erro de Variância: também conhecido como *overfitting*, ocorre quando o modelo não consegue generalizar para dados não vistos (conjunto de testes) e não acerta o atributo alvo corretamente. Geralmente ocorre quando o modelo de aprendizagem é mais complexo do que os padrões apresentados pelos dados. Para solucioná-lo, modelos mais simples devem ser utilizado, ou uma seleção/regularização de atributos deve ser realizada.

Uma das técnicas mais comumente utilizadas para estimar o desempenho de forma mais robusta é a estratégia *cross-validation*. Essa técnica consiste em: (I) dividir o conjunto de dados em K partições de forma randômica e mutualmente exclusivas com tamanhos aproximadamente iguais; (II) treinar e estar o modelo K vezes; (III) fazer cada iteração de treinamento com K-1 partições e validar sobre o conjunto restante, permitindo que cada instância seja usado no conjunto de treinamento K-1 vezes e de validação 1 vez; (IV) atribuir o valor da avaliação final pela média da avaliação de cada conjunto de validação. O principal parâmetro desse algoritmo é o valor de K, que representa a quantidade de partições que os dados serão divididos. O valor mais recomendado para o K é 10 (Kohavi 1995).

Outro importante parâmetro a ser definido para estimar o desempenho do modelo é se os dados deverão ser estratificados. A estratificação significa que os dados de treinamento e teste deverão ser amostrados de acordo com a proporção original de cada classe. Por exemplo, no problema binário desbalanceado, podemos ter uma relação entre a classe 1 e 0 de 2:8 (2 para 8), ou seja, para cada 10 instâncias, 2 são da classe 1 e 8 são da classe 0. Logo, para o exemplo citado, quando o parâmetro estratificado é ativado, o grupo de treinamento e o grupo de testes deverão ter em cada partição proporção de instância de classe 1 e 0 de 2:8.

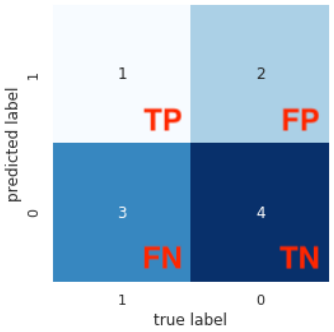
2.5.2 Métricas para estimar desempenho do modelo

Para estimar o desempenho de modelos de classificação, uma das técnicas mais usadas na literatura é a matriz de confusão (Ting 2017). A matriz de confusão relaciona o resultado da predição do modelo com a classe real, sendo que as colunas indicam os índices das classes verdadeira e as linhas indicam das classes preditas pelo modelo. A Figura 2.18 apresenta uma matriz de confusão para um exemplo com duas classes. Percebe-se

que o formato de apresentação da matriz de confusão pode mudar dependendo do eixo que representa a predição e a classe real. Nesta tese, como temos um problema de classificação binária (alunos evadidos e alunos persistentes), usaremos o padrão do pacote computacional de aprendizagem de máquinas *sklearn* da linguagem Python (Pedregosa et al. 2011), em que a Classe Negativa (Classe 0 na Figura 2.18) representa os alunos persistentes e a Classe Positiva representa os alunos evadidos (Classe 1 na Figura 2.18). Geralmente a "Classe 1"é o valor alvo de maior interesse no trabalho, no nosso caso, os alunos evadidos. Para o problema geral de classificação com n classes, teremos uma matriz de confusão nxn, sendo que o elemento (i, j) corresponde à quantidade de vezes que a classe j foi prevista como classe i. Assim, os elementos da diagonal principal da matriz de confusão correspondem os acertos do modelo em predizer corretamente as classes, enquanto que os elementos fora da diagonal principal correspondem aos erros de predição. Para o caso particular de classificação com duas classes, a matriz de confusão possui quatro elementos, que são descritos a seguir:

- ***True Positive* - (TP)**: Corresponde ao número de instâncias positivas classificadas corretamente localizadas. Nesta tese, são os alunos evadidos classificados de forma correta.
- ***True Negative* - (TN)**: Corresponde ao número de instâncias negativas classificadas corretamente. Nesta tese, são os alunos persistentes classificados de forma correta.
- ***False Negative* - (FN)**: Corresponde ao número de instâncias positivas classificadas incorretamente como negativas. Nesta tese, são alunos evadidos erroneamente classificados como persistentes e iremos considerar como o pior dos dois erros possíveis.
- ***False Positive* - (FP)**: Corresponde ao número de instâncias negativas classificadas incorretamente como positivas. Nesta tese, são alunos persistentes erroneamente classificados como evadidos.

Figura 2.18: Exemplo da matriz de confusão para o problema de classificação com duas classes.



Fonte: Elaborada pelo autor.

A partir da matriz de confusão, podem ser geradas várias métricas de avaliação de

desempenho do modelo treinado. As métricas mais utilizadas no problema de classificação são o *Recall* e a Precisão (*Precision*, em inglês). A Precisão do modelo é calculada usando a linha dos rótulos previstos como Classe Positiva (primeira linha da matriz de confusão). Essa métrica indica o quão bom é o nosso modelo sobre a predição positiva. Ou seja, dadas todas as instâncias que o modelo previu que são positivas, ela determina a porcentagem de acerto. Já a métrica *Recall* é calculada usando a coluna que contém os rótulos verdadeiros da Classe Positiva (primeira coluna da matriz de confusão). Essa métrica informa a taxa de acerto das instâncias que de fato são positivas e é importante quando queremos identificar o máximo de instâncias positivas da base de dados, como é o caso do problema abordado nesta tese, em que é mais importante identificar o aluno evadido frente ao aluno persistente.

As definições formais de *Recall* e Precisão são apresentadas na Tabela 2.3. É importante ressaltar que as métricas *Recall* e Precisão têm características assimétricas, pois seus cálculos dependem dos valores de FP e FN, respectivamente, sendo que esses não estão relacionados entre si e os seus valores podem ser diferentes de acordo com as definições para Classe 0 e Classe 1.

Uma maneira de mesclar as métricas Precisão e *Recall* é através da métrica F1 (Ting 2017), vide Tabela 2.3, que pode ser interpretada como uma média harmônica entre a Precisão e o *Recall*, com variação entre 1 (melhor avaliação) e 0 (pior avaliação). Ao contrário da média aritmética normal, que atribui o mesmo peso a todos os valores, a média harmônica atribui um peso maior a valores baixos. Isso significa que teremos uma pontuação F1 alta apenas se a Precisão e o *Recall* forem altas (Chicco & Jurman 2020). Outro importante detalhe é que, da mesma forma como acontece com a Precisão e *Recall*, o valor da F1 é assimétrico, pois muda de acordo como é definido a Classe 1 e a Classe 0.

Diferentemente das métricas anteriores, a métrica Acurácia Balanceada (AB), também conhecida como *Unweighted Average Recall* (UAR), vide Tabela 2.3, é uma métrica mais robusta para dados desbalanceados (He & Ma 2013). Essa métrica considera a média do *Recall* da classe negativa e positiva, não sendo afetada por uma alteração na frequência de ocorrência da classe (Schuller et al. 2009, Kaya & Karpov 2018). Outra vantagem é a simetria da métrica, ou seja, ela independe de que é definido como Classe 1 ou 0 entre a classe majoritária e minoritária, facilitando a análise dos dados.

Outra métrica bastante interessante é a *G-mean*, vide Tabela 2.3. A *G-mean* pode ser entendida em termos da geometria como o tamanho do lado de um quadrado cuja área é igual à área de um retângulo com lados do tamanho do *Recall* da classe positiva e negativa. Ela é recomendada quando o resultado é uma média normalizada entre números que não estão necessariamente nas mesmas escalas (Fleming & Wallace 1986).

A curva ROC (*Receiver Operating Characteristic*) (Kubat & Matwin 1997) é uma técnica bastante utilizada para avalizar o desempenho de modelos de classificador de duas classes. Nessa técnica, traça-se uma gráfico no qual o eixo da abcissa corresponde aos valores de taxas positivas verdadeiras (TPRs) [2] e o eixo da ordenada corresponde aos valores das taxas de falso positivo (FPRs) (Swets 1988). Então, varia-se um dos hiperparâmetros do modelo para se obter os pontos que compõem o gráfico da curva ROC. TPR e FPR são definidos pelas Equações 2.4 e 2.5, respectivamente.

[2] corresponde ao *recall*

Tabela 2.3: Métricas de Avaliação.

Metric	**Formula**
Precisão	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
AB	$\frac{TP}{2(TP+FN)} + \frac{TN}{2(TN+FP)}$
F1	$2 * \frac{Precision * Recall}{Precision + Recall}$
AUC	$\int_{x=0}^{1} \text{TPR}(\text{FPR}^{-1}(x))dx$
G-mean	$\sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Fonte: Elaborada pelo autor.

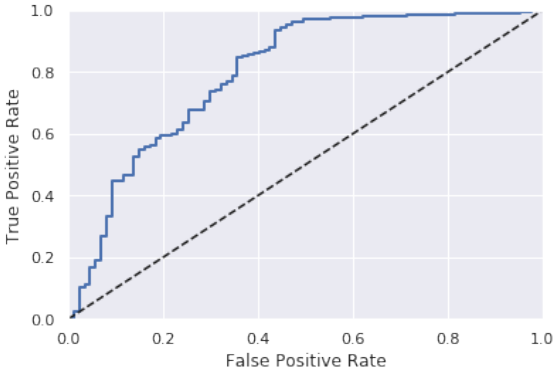
$$TPR = \frac{TP}{TP+FN} \tag{2.4}$$

$$FPR = \frac{FP}{TN+FP} \tag{2.5}$$

Ao avaliar modelos de classificação com várias taxas de erro, as curvas ROC são capazes de determinar qual proporção de instâncias será classificada corretamente para um determinado FPR. A Figura 2.19 apresenta uma exemplo de curva ROC para a matriz de confusão mostrada na Figura 2.18.

Enquanto as curvas ROC fornecem um método visual para determinar a eficácia de um classificador, a área sob a curva ROC (denominada de AUC, vide Tabela 2.3) se tornou a métrica padrão para avaliar classes desbalanceadas (Bradley 1997). Isso se deve ao fato de o valor calculado ser independente do limite selecionado e das probabilidades anteriores, gerando um número único que pode ser usado para comparar os classificadores. A AUC indica a probabilidade de que, se tomarmos quaisquer duas observações de nossas previsões, elas serão ordenadas da maneira correta. Por exemplo, dado o problema da classificação binária (negativo e positivo), um valor de AUC de 0.9 e a saída do modelo com valores de probabilidade (exemplo da regressão logística, sendo valores mais próximos de zero (0) a classe negativa e valores mais próximos de um (1) a classe positiva), em 90% dos casos, as instâncias previstas como classe positiva possuem valores de saída do modelo maiores que as instâncias previstas como classe negativa (Google 2020). É

Figura 2.19: Exemplo de curva ROC.



Fonte: Elaborado pelo autor.

importante enfatizar que, tal como a Acurácia Balanceada, as métricas *G-mean* e AUC são simétricas

Por fim, outra métrica robusta ao desbalanceamento de dados é a Correlação de Matthews (MCC). Essa métrica trata a classe verdadeira e a classe prevista como duas variáveis (binárias) e calcula seu coeficiente de correlação (de maneira semelhante ao cálculo do coeficiente de correlação entre as duas variáveis). Quanto maior a correlação entre os valores verdadeiros e os previstos, melhor a predição. Quando o classificador é perfeito (FP = FN = 0), o valor de MCC é 1, indicando uma correlação positiva perfeita. A MCC leva em consideração todos os quatro valores na matriz de confusão, e um valor alto (próximo a 1) significa que ambas as classes são bem previstas, mesmo sobre classes desbalanceadas. Ele varia no intervalo de -1 e $+1$, alcançados em caso de classificação incorreta e classificação perfeitas, respectivamente, enquanto MCC = 0 é o valor esperado para o classificador de lançamento de moedas (Chicco & Jurman 2020).

Na Tabela 2.4, estão sumarizadas as características das métricas de avaliação dos modelos apresentadas nessa seção.

Tabela 2.4: Comparação entre métricas de avaliação de modelo.

Nome	**Descrição**
Precisão	Dadas todas as instâncias que o modelo previu de uma classe, indica a taxa de acerto dessa predição. Apresenta valores diferentes de acordo com a classe (assimetria).
Recall	Indica a taxa de acerto das instâncias que de fato pertencem a uma dada classe. Apresenta valores diferentes de acordo com a classe (assimetria). Para o problema de classes desbalanceadas, é uma indicação interessante para verificar a taxa de acerto da classe minoritária.
F1	Média harmônica entre Precisão e *Recall*, ou seja, apenas apresenta valores altos se ambas possuem valores altos. Apresenta valores diferentes de acordo com a classe (assimetria)

AB	Representa a média do *Recall* da classe positiva e negativa, o que a torna simétrica, ou seja, independente ao que é definido como classe positiva ou negativa. Essa média é robusta para dados desbalanceados.
G-mean	Média geométrica entre o *Recall* da classe positiva e negativa. Recomendada quando o resultado é uma média normalizada entre números não necessariamente nas mesmas escalas. É uma média simétrica e robusta para dados desbalanceados.
AUC	Indica a probabilidade de que, se tomarmos quaisquer duas observações de nossas previsões, elas serão ordenadas da maneira correta. É uma métrica robusta aos dados desbalanceados e simétrica, sendo uma das mais utilizadas no problema de classes desbalanceadas.
MCC	Métrica que indica a correlação entre a classe verdadeira e a classe prevista. Quanto maior a correlação entre os valores verdadeiros e previstos, maior o valor da métrica. Seus valores variam de -1 a 1 e são robustos ao desbalanceamento, além de ser simétrica.

Fonte: Elaborada pelo autor.

2.5.3 Comparação entre os modelos treinados

Em relação à comparação entre os modelos de aprendizagem, atualmente os t-testes pareados estão dando lugar ao teste de McNemar e 5X2 *cross-validation* (Demšar 2006).

O teste de McNemar (McNemar 1947) é um teste estatístico não paramétrico para comparações pareadas. Ele é aplicado a dados nominais emparelhados com base em uma versão da matriz de confusão 2x2 (às vezes também chamada de tabela de contingência 2x2) que compara as predições de dois modelos entre si, como visto na Figura 2.20 (Raschka 2018).

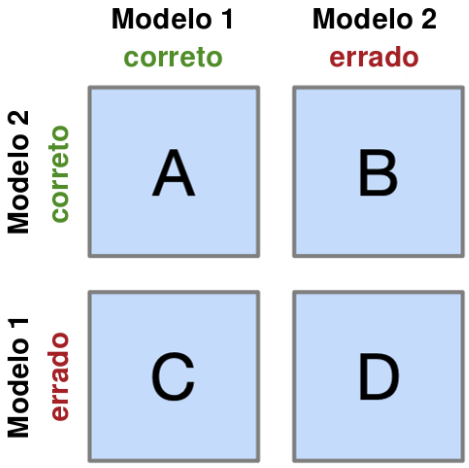
O teste de McNemar assume como hipótese nula que os modelos têm o mesmo desempenho a partir do cálculo do teste do chi-quadrado, dado pela seguinte fórmula:

$$X^2 = \frac{(B-C)^2}{B+C} \tag{2.6}$$

Após o cálculo, geralmente assume-se que o valor de alfa (também conhecido como significância) é igual a 0,05 e se o p-valor for menor que o nível de significância escolhido, podemos rejeitar a hipótese nula de que os desempenhos dos dois modelos são iguais. É importante destacar que o teste de McNemar assume números relativamente grandes nas células B e C (por exemplo, maiores que 25).

O t-teste pareado 5x2CV é um procedimento para comparar o desempenho de dois modelos de aprendizagem, proposto para resolver deficiências de outros métodos como o t-teste pareado com reamostragem e o t-teste pareado com *cross-validation* (Dietterich 1998). Como no teste McNemar, aqui assume-se como hipótese nula que os modelos têm o mesmo desempenho. Então, a divisão é repetida (50% de treinamento e 50% de dados de teste) cinco vezes. Em cada uma das 5 iterações, os classificadores C1 e C2 são ajustados e seus desempenhos são avaliados. Em seguida, os conjuntos de treinamento e

Figura 2.20: Matriz de confusão no contexto do teste McNemar



Fonte: Adaptada de Raschka (2018)

teste são mudados (o conjunto de treinamento se torna o conjunto de teste e vice-versa) e calcula-se o desempenho novamente (Raschka 2018). Após esse procedimento, o valor da estatística t é calculado pela seguinte fórmula:

$$t = \frac{ACC}{\sqrt{(1/5)\sum_{i=1}^{5} s_i^2}} \tag{2.7}$$

Usando a estatística t, geralmente assume-se que o valor de alfa é igual a 0,05 e se o p-valor for menor que o nível de significância escolhido, podemos rejeitar o hipótese nula de que os desempenhos dos dois modelos são iguais.

2.6 Interpretação de modelos

A Interpretabilidade de dados (também conhecido como *Explainable* AI em inglês - XAI) pode ser definida como o grau em que um humano pode entender a causa da tomada de decisão de um modelo baseado em dados (Miller 2017). Há modelos baseados em técnicas de aprendizagem de máquinas cujas estruturas são mais fáceis de interpretação por humanos, como as árvores de decisão, regressão logística e regressores lineares. Essas abordagens são denominadas de caixas-brancas. Por outro lado, modelos cujas suas estruturas são de difícil interpretação por humanos são chamados de caixas-pretas, como são os casos de técnicas das redes neurais artificiais (exemplos mais conhecido são as redes neurais do tipo *Deep Learning*), *Suport Vector Machine* (SVM) e *Random Frorests* (RF).

Esse tema de XAI vem ganhando força a partir da necessidade que as instituições tem em responder as seguintes perguntas: "como posso confiar nesse seu modelo?", "como

seu modelo toma suas decisões?"(Ribeiro et al. 2016, Mori & Uchihira 2018, Johansson et al. 2011). Então é importante se ter em mente que entender como os modelos baseados em dados funcionam é fundamental para identificar e evitar possíveis vieses contidos nos dados (até mesmo preconceituosos), tornando os modelos mais justos (Corbett-Davies & Goel 2018).

Como já apresentado na Figura 2.16, os modelos de regressão logística e árvores de decisão são os que têm melhor interpretação.

Os modelos baseados em regressão logística podem ser facilmente interpretados a partir dos coeficientes gerados após o seu treinamento. O valor dos coeficientes estão relacionados com os "odds" (probabilidade de evento dividido pela probabilidade de nenhum evento) de uma classe para cada atributo de acordo com as Equações 2.8 e 2.9 (Molnar 2019).

$$log\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.8}$$

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right) \tag{2.9}$$

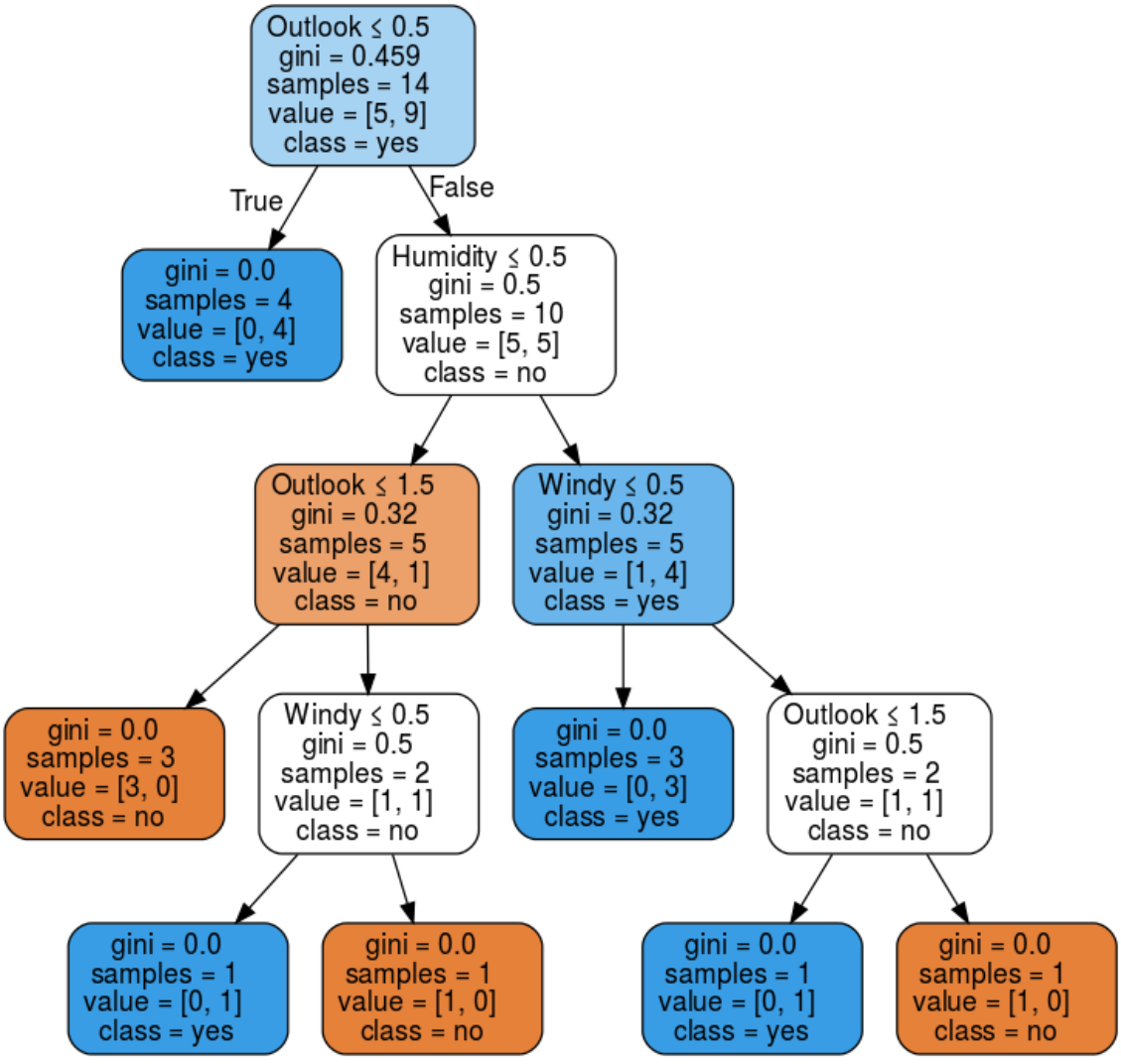
Por exemplo, se há um peso (β_p) de 0,7 para um dado atributo, então a cada aumento do seu valor em uma unidade, dobra-se a chance daquela instância ser de uma classe 1 e não da classe 0, mantendo-se todos os outros atributos constante. A partir desses coeficientes podemos, então, extrair a informação da contribuição de cada atributo na probabilidade de uma dada classe.

Para as árvores de decisão, a técnica mais utilizada para extração de conhecimento é a *WHAT-IF Rule*. Essa técnica é caracterizada como um conjunto de regras do tipo IF-THEN, que mostra a tomada de decisão realizada por um modelo de aprendizagem baseado em árvore de decisão. A interpretação é simples: para cada nó da árvore (começando pelo nó raiz) adiciona-se o "AND"e a condição descrita no nó e, então, segue-se para os próximos. Ao atingir o próximo nó, repete-se o mesmo processo de forma recursiva até alcançar as folhas da árvore, quando a conclusão prevista do "THEN"é informada. Por exemplo, na Figura 2.21 é apresentada a visualização de uma DT para o famoso problema do *Play Golf*. Uma regra extraída desse modelo é a seguinte:

IF Outlook > 0.5 AND Humidity > 0.5 AND Windy < 0.5 THEN Class = yes

A grande desvantagem para esse método é que pequenas mudanças na base de dados podem gerar árvores de decisão diferentes e, consequentemente, regras diferentes (Molnar 2019). Como principal vantagem, um usuário não especialista em Ciência de Dados pode usar diretamente o modelo treinado para detectar as causas do problema em análise e tomar decisões pertinentes (Márquez-Vera et al. 2013).

Figura 2.21: Exemplo de resultado da técnica DT no *beanchmark Play Golf*.



Fonte: Elaborada pelo autor.

Capítulo 3

Trabalhos relacionados

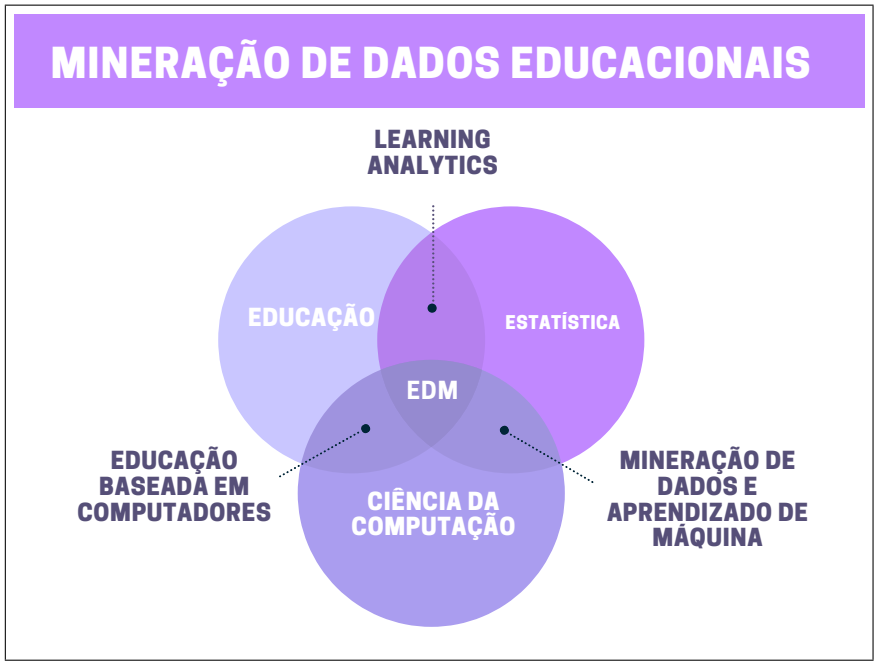
No contexto de dados educacionais, as duas principais áreas emergentes baseadas em Ciência de Dados são a Analítica de Aprendizagem (conhecida como *Learning Analytics* em inglês - LA) e a Mineração de Dados Educacionais (conhecida como *Educational Data Mining* em inglês - EDM).

A primeira (e bastante popular) conferência científica sobre *Learning Analytics* (LAK 2011) define LA como sendo "a medição, coleta, análise e relatório de dados sobre os alunos e seus contextos, com o objetivo de entender e otimizar aprendizagem e os ambientes em que ocorre ". Assim, de acordo com essa definição, o foco está principalmente em otimizar a aprendizagem do aluno, como criar um modelo desse aluno e suas interações com o ambiente de aprendizagem. Já a Mineração de Dados Educacionais é definida como a interseção entre as grandes áreas de estatística, mineração de dados e educação (Figura 3.1), a fim de analisar e extrair novos padrões e conhecimento sobre esses dados (Romero & Ventura 2010). Com isto, o objetivo da EDM é aplicar as ferramentas de mineração de dados no contexto de educação, sem necessariamente com focar em modelar o aluno. Nesta tese, o problema da evasão escolar será abordado a partir da visão da EDM.

Apesar das pequenas diferenças conceituais, os modelos de predição de evasão e desempenho de estudantes são uma das atividades mais antigas e estudadas em ambas as áreas de EDM e LA (Romero & Ventura 2019*b*). A relevância da predição de evasão escolar se dá pelo fato que quanto mais cedo for detectada uma possível desistência ou insucesso na atividade acadêmica, é possível intervir e evitar a continuação de um baixo desempenho ou até a evasão do aluno (Delen 2011), (Thammasiri et al. 2014), (Burgos et al. 2018), (Nelson et al. 2012), (Jayaprakash et al. 2014), (Romero & Ventura 2019*b*). Além disso, quando comparada com os métodos estatísticos tradicionais, essa abordagem gera modelos com menos restrições (por exemplo, normalidade, independência, colinearidade, etc.) e que são capazes de produzirem melhores predições (Thammasiri et al. 2014).

De forma mais detalhada, a predição de desempenho escolar do aluno tem como objetivo prever o valor da variável que representa o seu desempenho, sendo uma variável do tipo numérica (por exemplo, a nota do aluno em uma determinada disciplina) ou categórica (por exemplo, o status escolar do aluno: evadido ou regular). Quando a previsão envolve variáveis numéricas, geralmente são utilizadas técnicas de regressão analítica, as quais encontram a relação entre uma variável dependente e uma ou mais variáveis independentes. Já quando a variável em questão é representada por valores categóricos, geralmente são utilizadas técnicas de classificação ou agrupamento, em que itens individuais

Figura 3.1: EDM representada como uma área interdisciplinar.



Fonte: Elaborada pelo autor.

são colocados em grupos com base em cálculos de similaridade entre as suas instâncias.

É importante ressaltar que a qualidade dos dados a ser utilizada para a construção do modelo é um fator determinante para a obtenção de bons resultados. No contexto de predição de evasão escolar, há duas características comuns que podem comprometer a qualidade na base de dados e tornar o processo de extração de conhecimento mais complexo: alta dimensionalidade dos dados (muitas variáveis independentes) e dados desbalanceados. O primeiro problema ocorre pela grande quantidade de registros de atributos (em inglês, *features*) sobre os alunos nos sistemas de controle acadêmicos, ambientes virtuais de aprendizagem e outros sistemas digitais de monitoramento. A fim de diminuir essa alta dimensionalidade, geralmente utiliza-se técnicas de seleção de atributos (Bolón-Canedo et al. 2016). Já o segundo problema é ocasionado pelo quantitativo significativamente menor de alunos que evadem quando comparado aos que persistem no curso, criando um problema de classificação binária desbalanceada. Para resolver esse problema, é necessário utilizar técnicas de balanceamento de dados em conjunto com métricas adequadas de avaliação do modelo, a fim de evitar o problema conhecido como Paradoxo da Acurácia, fenômeno que pode mascarar os resultados obtidos (Zhu & Davidson 2007).

Para posicionar este trabalho no estado da arte de modelos de predição de evasão escolar baseada em dados, foi realizada uma revisão sistemática da literatura na base de dados *Scopus*, na data de 02/05/2019, seguindo o protocolo da Tabela 3.1.

A busca definida retornou 107 trabalhos, dos quais 12 atenderam os critérios selecionados. Esses trabalhos estão resumidos na Tabela 3.2, em que são destacados: as técnicas

Tabela 3.1: Protocolo de revisão sistemática da literatura.

Parâmetro	**Descrição**
Questões de pesquisa	Quais são as técnicas de: seleção de atributos, balanceamento, modelo de aprendizagem, métrica de avaliação de modelo e interpretação do modelo treinado
Critérios de Inclusão	Artigos que propõem/relatam modelo preditivo de evasão sobre dados reais e desbalanceados
Critérios de Exclusão	Artigos que: não apresentam algum método, técnica, modelo ou ferramenta para predição de evasão escolar; não apresentam as métricas utilizadas para avaliação de desempenho do modelo preditivo; não apresentam resultados extraídos de um contexto real de ensino
String de busca	TITLE-ABS-KEY (("dropout"OR "drop out"OR "drop-out"OR "freshmen"OR "retention"OR "at-risk"OR "Withdrawal"OR "fail"OR "Attrition") AND ("predict"OR "predicting"OR "prediction"OR "predictive"OR "model"OR "Classification"OR "detection"OR "forecasting") AND ("education"OR "academic"OR "school"OR "student") AND ("imbalanced"OR "disproportion"OR "asymmetry"OR "unbalance"))

Fonte: Elaborada pelo autor.

de seleção de atributos, os métodos de balanceamento, os modelos de aprendizagem, as métricas de avaliação de modelo e as técnicas de interpretação e extração de conhecimento do modelo treinado.

Tabela 3.2: Revisão sistemática da literatura.

Referencia	**Sel. de Atributos**	**Balanceamento**	**Modelo**	**Avaliação**	**Interpretação**
(Al-Jallad et al. 2019)	Utiliza, mas não descreveu	SMOTE	RG method (based on Induction and DT	Especificidade, Recall, Acurácia	WHAT-IF rule
(Delen et al. 2019)	Elastic net	SMOTE	Bayesian Belief Network-Driven Probabilistic	AUC, Acurácia, Recall, Especificidade, G-Mean	Análise de Sensibilidade, WHAT-IF rule
(Ramentol et al. 2019)	-	6 variações do SMOTE, SPIDER2, Cost-Sensitive	Probabilistic Rough Set	AUC	-
(Nadar & Kamatchi 2018)	-	Cost-Sensitive	Não-Linear SVM	ROC, F-Score, and Precisão and Recall	-
(Periwal & Rana 2017)	-	-	NB, KNN, DT, LR	Recall, Especificidade, FPR, FNR, Acurácia	-
(Hlosta et al. 2017)	-	Class Weights	LR, SVM, RF, NB, and XGBoost.	Precision-RecallCcurve	Análise de sensibilidade
(Punlumjeak et al. 2017)	Mutual Information	Downsample	Ensemble (DT, RNA), RF, Adaboost (DT, ANN)	Acurácia, Precisão, Recall	-
(Márquez-Vera et al. 2016)	3 attribute subset and 7 single attribute evaluators	SMOTE, Cost-Sensitive	Interpretable Classification Rule Mining, NB, SVM, KNN, Classification Rules, DT	Recall, Especificidade, G-Mean, AUC	WHAT-IF rule
(Ram et al. 2015)	-	SMOTE, SplitBal with Ensemble	SVM, DT, NB, RF	AUROC, Precisão, Recall and F-Score, Acurácia	-

(Thammasiri et al. 2014)	-	Oversample, Downsample and SMOTE	LR, DT, ANN, SVM	Recall, Precision, Accuracy, F-Score, FPR, G-Mean, CorrelationCoe	DTOR model Análise de Sensibilidade
(Márquez-Vera et al. 2013)	3 attribute subset and 7 single attribute evaluators	SMOTE	Genetic Programming, DT, Rule Induction	Acurácia, Recall, Especificidade, G-Mean	WHAT-IF rule
(Ghanem et al. 2008)	-	-	Probabilistic Relational Models	AUC	-

Fonte: Elaborada pelo autor.

São sinônimos os termos: *Recall, Sensitivity* e TPR; Especificidade e TNR. Siglas: *Logistic Regression* (LR), *Support Vector Machines* (SVM), *Random Forest* (RF), *Naive Bayes* (NB), Árvore de Decisão (DT), *K-Nearest Neighbor* (KNN), Redes Neurais Artificiais (RNA)

A primeira conclusão é a complexidade de afirmar a superioridade de um modelo sobre o outro, devido às diferentes métricas de desempenho e aos diferentes contextos institucionais que precisam ser levados em conta. Outros destaques são: que a minoria dos estudos (5 de 12) realizaram seleção de atributos a fim de amenizar o problema da alta dimensionalidade; a maioria dos trabalhos (7 vezes) utilizaram a técnica SMOTE para balanceamento dos dados; há uma grande variedade nos modelos de aprendizagem, entretanto as técnicas DT (7 vezes), SVM (5 vezes), NB (4 vezes), LR (3 vezes), RF (3 vezes) apareceram ao menos em 3 trabalhos distintos; em relação as métricas de avaliação, temos com maior frequência a Recall (9 vezes), Acurácia (6 vezes), Especificidade (5 vezes), G-Mean (4 vezes), AUC (3 vezes), F-Score (3 vezes), ressaltando também mais de uma métrica pode ter sido utilizadas em cada estudo; no tocante à interpretação dos modelos treinados, apenas 6 trabalhos utilizaram alguma técnica, sendo as mais recorrentes WHAT-IF rule (4 vezes) e análise de sensibilidade (3 vezes).

Além dos trabalhos levantados pela RSL, outros artigos foram analisados pois envolviam a temática de predição de evasão ou performance do aluno. A Tabela 3.3 apresenta esses trabalhos relacionados, destacando os modelos de aprendizagem utilizados para predição, as variáveis independentes e a métrica de avaliação.

Tabela 3.3: Comparação com trabalhos relacionados.

Referência	**Modelo de aprendizagem**	**Variáveis independente**	**Avaliação**
(Huang & Fang 2013)	Regressão linear, MLP, RBF, SVM	Performance acadêmica	Acurácia
(Rovira et al. 2017)	Regressão logística, NB, SVM, RF, Adaptive Boosting (AB), filtro colaborativo, regressão linear	performance acadêmica	acurácia, recall, precisão, F1
(Li et al. 2013)	Análise de componentes principais	performance acadêmica	análise do PCA
(Xu et al. 2017)	Fatoração de matriz probabilística	Performance acadêmica, dados demográficos	Erro médio quadrático
(Meier et al. 2015)	Algoritmo próprio utilizando cálculo de semelhança e regressão	Performance acadêmica	Erro médio absoluto
(L. G. F. Silva & Fagundes 2017)	Regressão linear, Stepwise	Dados demográficos do aluno e do curso	Erro médio absoluto e o erro absoluto relativo.
(Y. Amaya & Heredia 2015)	DT	Performance acadêmica, dados socioeconômicos	Não foi possível identificar

(Shahiri et al. 2015)	DT, RNA, NB, K-NN e SVM	Performance acadêmica, dados demográficos, atividades extra-curricular, rede de interação social	Acurácia
(Burgos et al. 2018)	Regressão logística	Performance acadêmica	Acurácia, matriz de confusão
(Lykourentzou et al. 2009)	Combinação de MLP, SVM e Ensemble	Dados demográficos, Performance acadêmica	Acurácia, sensitividade e precisão
(Asif et al. 2017)	DT, Agrupamento hierárquico	Performance acadêmico	Acurácia, Kappa, Matriz de Confusão
(Gray & Perkins 2019)	DT	Engajamento, curso, escola, ano de estudo	TP, FP, Precisão, AUC

Fonte: Elaborada pelo autor.

A grande maioria dos trabalhos (11 de 13) utiliza da performance acadêmica como variável independente, sendo na maioria dos casos o único tipo de variável utilizada (7 vezes). Em apenas 5 vezes, os dados demográficos foram utilizados também como entrada para o modelo. Em 6 trabalhos, a acurácia foi utilizada como métrica de avaliação. Em nenhum deles foram utilizados métricas mais robustas como G-mean, AB, AUC ou MCC. É importante lembrar que, em alguns trabalhos, o objetivo era prever a performance e não a classificação do aluno entre evadido ou persistente.

A grande crítica observada nesses trabalhos é não abordar a questão como um problema de classes desbalanceados e os cuidados que devem ser tomados a partir dessa visão. Logo, o uso da métrica de acurácia para avaliar os modelos de aprendizagem pode tê-los levado a caírem no problema do paradoxo da acurácia, mascarando os resultados obtidos. Alguns trabalhos utilizaram métricas um pouco mais robustas como F1, Recall e Precisão, entretanto, como visto na seção teórica, os valores das métricas não são simétricos, logo, essas métricas estão sujeitas de como é definida a matriz de confusão. Além disso, os trabalhos estão geralmente focados nos modelos de aprendizagem para predição da performance do aluno ou evasão do aluno. Nenhum deles enfatiza uma visão mais genérica como o processo que deveria ser utilizado.

Por fim, foi observado que a falta de consenso sobre quais técnicas e algoritmos devem ser adotados nas diversas atividades associadas ao problema de predição de evasão baseada em dados deve-se à complexidade do problema. Assim, torna-se clara a necessidade de se estabelecer um processo padronizado para modelar o fluxo que vai desde a obtenção de informação sobre evasão escolar, o qual inclua atividades da captura de dados brutos, até a disponibilização de informação em formato útil a gestores e professores. Dessa forma, habilita-se uma análise de desempenho dos diversos algoritmos e técnicas associados com cada atividades desse processo. Neste sentido, na próxima seção, são apresentados as etapas adotadas para modelar o processo baseado em dados para obtenção de informação sobre evasão escolar e a fundamentação conceitual dos algoritmos, cujos desempenho serão investigados, com indicação em quais etapas estes algoritmos

devem ser empregados.

Capítulo 4

Metodologia

O problema da evasão escolar, como já exposto na introdução, tem um caráter complexo e contextualizado frente à instituição de educação que sofre desse fenômeno. Portanto, os fatores que levam o aluno a evadir podem mudar de acordo com a realidade local, fatores socioeconômicos, organizacionais, motivação individual, entre outros. Diante desse cenário, criar uma máquina de aprendizado universal para predição de evasão escolar se torna uma tarefa complexa que, provavelmente, não irá alcançar os resultados com precisão aceitável para subsidiar um gestor ou professor na sua tomada de decisão. Logo, nesta tese, propomos um processo sistemático a fim de modelar um fluxo de informação desde a definição do problema até a geração de informação útil a gestores e professores. O processo é composto pelas seguintes etapas: "Entender o problema", "Entender os dados", "Engenharia de atributos"(mais conhecida por *Feature Engineering*, em inglês), "Seleção de atributos"(mais conhecida por *Feature Selection*, em inglês), "Balanceamento de dados", "Modelos", "Avaliação"e "Interpretação".

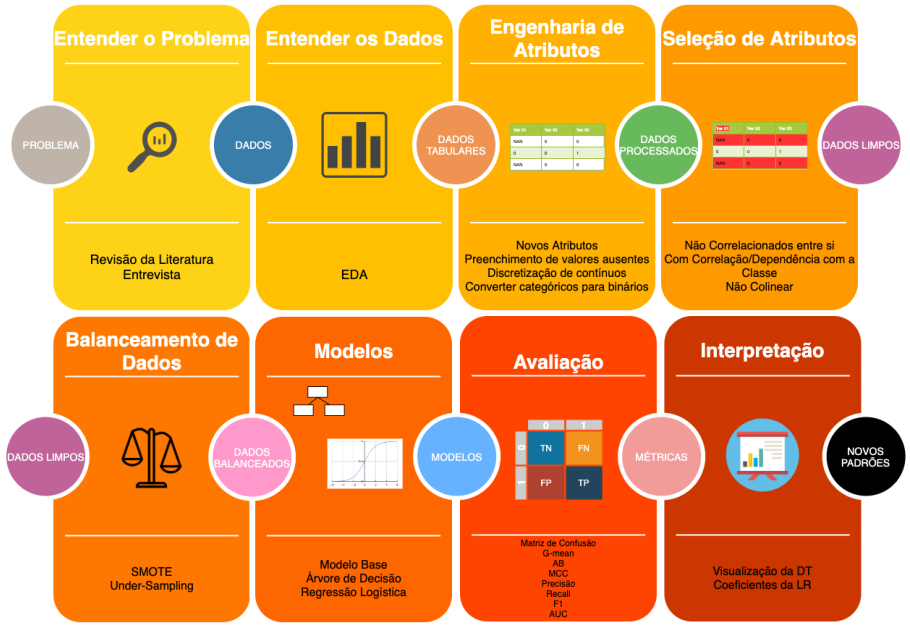
A partir do processo sistemático definido aqui nesta tese, será verificada a seguinte hipótese: "o modelo preditivo de evasão escolar gerado tem um desempenho melhor frente à suposição que o aluno irá evadir caso: (I) seu desempenho escolar médio e (II) sua frequência média estejam inferiores ao desempenho mínimo estabelecido pela organização didática?".

De forma resumida, é apresentada na Figura 4.1 a sistematização do processo proposto desta tese para criação do modelo preditivo e extração de conhecimento, sendo indicadas quais técnicas de Ciência de Dados deverão ser utilizadas em cada uma das etapas. Ressalta-se que todas as técnicas de Ciência de Dados a serem empregadas na nossa propostas foram apresentadas em detalhes no Capítulo 2 desta tese.

É importante ressaltar que cada uma das etapas que compõe o nosso processo sistemático para geração de modelos de predição de evasão escolar foi concebida para buscar:

1. Modelos e técnicas com compreensão simples por humanos, a fim de auxiliar a tomada de decisão por gestores e professores, evitando, assim, modelos considerados caixas-pretas.
2. Modelos e técnicas que permitam a visualização de seus resultados.
3. Modelos e técnicas que enfatizem o problema de evasão escolar, pois consideramos nesta tese que é mais importante identificar o aluno o qual potencialmente possa evadir à identificação do aluno persistente.

Figura 4.1: Fluxo de extração de conhecimento para modelos de predição de evasão.



Fonte: Elaborada pelo autor.

A seguir, serão descritas as atividades em cada uma das etapas do Processo Sistemático proposto.

4.1 Entender o problema

Uma célebre frase, por vezes associada ao Filósofo John Dewey (Dewey 1938) e por outras a Charles Kettering (Líder de Pesquisa na General Motors), sintetiza a importância dessa etapa: "Um problema bem definido está metade resolvido".

Antes de apresentar os passos propostos nessa etapa, é importante destacar algumas características da mente humana que dizem respeito sobre a vulnerabilidade a ilusões e falácias. Uma vez que nossos cérebros são "limitados em sua capacidade de processar informações e evoluíram em um mundo sem ciência e outras formas de checagem de fatos"(Pinker & Motta 2018), desenvolvemos vieses cognitivos, como exemplo: "raciocínio motivado (dirigir um argumento para uma conclusão preferida, em vez de segui-lo para onde ele conduz), avaliação tendenciosa (criticar a prova que não confirma uma posição preferida e aprovar indícios que a sustentam) e um viés para o "meu lado"(Kahneman & de Arantes Leite 2012, Pinker & Motta 2018).

Diante desse cenários de vieses cognitivos, como primeiro passo, é sugerido pesquisar as abordagens já utilizadas sobre o problema, por exemplo, através da Revisão Sistemática da Literatura (RSL) (Kitchenham & Charters 2007). O objetivo é encontrar ideais, fatos e taxas-base que possam evitar os vieses cognitivos. Para isso, é proposto para a RSL a

identificação dos algoritmos e modelos utilizados para "Seleção de atributos", "Balanceamento", "Modelos de aprendizagem", "Métricas de avaliação"e "Interpretação"do modelo treinado. Além disso, é preciso identificar quais as principais variáveis independente utilizadas nas abordagens propostas. Essa etapa está sistematizada no Capítulo 3 desta tese.

Após o entendimento de como outros pesquisadores abordam o problema, uma sugestão de segundo passo é a realização da análise de contexto, seja a partir de entrevistas estruturadas ou informais aos interessados e/ou afetados pela solução (conhecidos como *stalkholders*). O objetivo aqui é relacionar as propostas da Revisão de Literatura com o contexto específico do estudo de caso. Algumas técnicas para realizar a análise de contexto podem ser baseadas em *Design* Instrucional ou em *Design Thinking* (Filatro 2019). Assim, com o intuito de melhor compreender o problema, a seguir são propostas perguntas importantes para entender o contexto da evasão escolar em uma dada instituição:

1. Qual o perfil do aluno com mais chance de evadir?
2. Quais dados sobre os alunos estão disponíveis e suas restrições?
3. Há alguma política de intervenção de permanência do aluno na instituição escolar?

Uma vez entendido o contexto, o terceiro passo sugerido é definir exatamente o que constituiria uma solução para esse problema (Cady 2017) e os princípios norteadores para uma boa solução. Por exemplo, que critérios constituem um projeto concluído e o que seria necessário para tornar o projeto um sucesso? Como princípio norteador muito utilizado na área de aprendizado de máquina, temos a navalha de Occam (*Occam's razor*, em inglês), que também é conhecido como princípio da parcimônia. Esse princípio coloca que dadas duas explicações equivalentes dos dados, a mais simples é preferível (Shavlik et al. 1990). Logo, quanto menor for a quantidade de parâmetros do modelo, menor será sua complexidade, o que é preferível.

Nesta tese, o critério de sucesso estabelecido é que o modelo gerado deve ser o mais interpretável possível e deve ser estatisticamente superior à heurística comumente citada por professores, que o aluno com potencial de evasão é aquele com:

- desempenho escolar médio **E** frequência média inferiores ao desempenho mínimo estabelecido pela organização didática.

Nesse caso, os valores considerados são nota mínima 60 e frequência mínima de 80%. Nesta tese, a heurística definida acima será chamada de modelo *Baseline*.

Com o problema definido, o quarto passo compreende as seguintes perguntas, mais específicas, que dizem mais respeito à área de ciência de dados:

1. Este é um problema de aprendizado supervisionado, não-supervisionado ou por reforço?
2. Se o aprendizado é supervisionado, é um problema de classificação ou regressão?
3. Se de for de classificação, é um problema binário ou de multi-classes?
4. Como será a validação do modelo?
5. Quão importante é a interpretação do modelo frente à sua acurácia?

Nesta tese, propomos que o problema de evasão escolar deve ser abordado por técnicas de aprendizado supervisionado. Mais especificamente, seria um problema de classificação

binária de classes desbalanceadas. Além disso, é importante que o modelo de aprendizagem deva ser facilmente interpretável, a ser o mais útil possível à tomada de decisão e à compreensão das relações das variáveis de entrada do modelo. O modelo também deve atingir um desempenho satisfatório no conjunto de testes (acima de 0.8 em ao menos três (3) entre G-mean, AUC, AB e MCC).

Na Figura 4.2, é sistematizada a sequência de passos sugeridas na Etapa de "Entender o Problema".

Figura 4.2: Sistematização da Etapa "Entender Problema".

Fonte: Elaborada pelo autor.

4.2 Entender os dados

Nas bases de dados educacionais, há dois principais grupos de atributos: os estáticos e os dinâmicos (Romero & Ventura 2019*a*). Os atributos estáticos são referentes ao perfil demográfico e socioeconômico, geralmente colhidos no ingresso do aluno na instituição, e registros educacionais anteriores ao ingresso, de tal forma que não são atualizados frequentemente. Já os atributos dinâmicos são referentes ao desempenho escolar do aluno (notas e frequência, por exemplo) ao longo do curso e o comportamento *online* relacionado aos cursos na modalidade de Educação a Distância (EaD), como dados relacionados à interação no Ambiente Virtual de Aprendizagem (AVA), o tempo na plataforma assistindo aos vídeos, quantidade de cliques, frequência de acesso ao material didático e quantidade de postagens em fóruns de discussão, dentre outros. Algumas instituições já

conseguem coletar dados a partir do celular dos alunos (Wang et al. 2018), como localização, hábitos e interações sociais. À medida que novas tecnologias ubíquas se tornam uma realidade, novos atributos poderão ser colhidos em continuamente, trazendo uma melhor compreensão do perfil dos alunos.

Nesta tese, usaremos atributos estáticos (referentes as condições socio-econômicas e demográficas) e atributos dinâmicos (referentes ao desempenho escolar e frequência do aluno).

Uma vez definido o problema pela ótica da Ciência de Dados, é fundamental o levantamento de quais dados sobre o aluno estão disponíveis. Em se tratando de ambiente escolar digitalizado, geralmente os dados estarão estruturados em banco de dados relacionais. Sendo essa a realidade, a pesquisa junto ao administrador do banco de dados da instituição é fundamental, a fim de responder as seguintes perguntas:

1. Quais as fontes dos dados (banco de dados, sensores, web, rede móvel) sobre os alunos estão disponíveis.
2. Qual será o formato em que os dados serão fornecidos para o modelo (xls, json, sql).
3. Qual a disponibilidade dos dados (o tempo de latência de atualização dos dados).
4. Quais restrições legais e institucionais de acesso a esses dados.

No estudo de caso (Capítulo 5) para validação do processo sistemático proposto nesta tese, utilizaremos um cenário de educação na modalidade presencial, por isso não serão utilizados registros de interação no AVA. Apesar disso, a nossa proposta é genérica e também pode ser utilizada para educação na modalidade de EaD.

Em geral, os dados educacionais institucionais estão armazenados em banco de dados relacionais, contendo diversas tabelas. Então, é fundamental que se integrem todos os atributos a serem utilizados nas análise em uma única tabela, que na área de Ciência de Dados é conhecida como tabela *flat*.

Uma vez criada a tabela unificada com os dados dos atributos a serem considerados nas análises, é realizado o processo de análise exploratório de dados a fim de compreender melhor a natureza e distribuição dos dados. Como foi visto no Capítulo 2, a EDA utiliza-se fortemente de ferramentas estatísticas e técnicas de visualização. Abaixo, seguem as técnicas que propomos para serem utilizadas nessa etapa, com respectivos propósitos:

1. Visualização da dimensionalidade dos dados: verificar a relação entre quantidade de instâncias e atributos e se o problema pode ser abordado como Big Data.
2. Visualização dos tipos de variáveis/atributos: verificar se os dados são numéricos, categóricos ou verdadeiro/falso. As técnicas estatísticas dependente do tipo do dado.
3. Levantamento dos dados ausentes: verificar a qualidade dos atributos e se devem ser utilizadas técnicas de imputação ou a remoção de atributo.
4. Analisar o balanceamento dos dados: verificar a proporção entre as classes de alunos evadidos (Classe 1) e alunos persistentes (Classe 0).
5. Visualização de correlação e dependência: verificar atributos correlatos (quando numéricos) e dependentes (quando categóricos).

6. Visualização da distribuição e medidas de estatísticas descritivas: visualizar as medidas e distribuições estatísticas dados a partir de gráficos como Violino, Box-plot, Q-Q e Histogramas.
7. Visualização de dados em baixa dimensão: uso do T-SNE para verificar se há padrões dos dados em baixa dimensionalidade.

É importante ressaltar que o fluxo de descoberta do conhecimento é iterativo, ou seja, podem ser necessárias várias iterações da EDA ao longo de todo o processo, a fim de se chegar em um entendimento satisfatório dos dados.

4.3 Engenharia de atributos

De posse da compreensão levantada pela análise exploratória dos dados, nesta etapa são realizadas as seguintes atividades sobre os dados: criação de novos atributos, preenchimento de dados ausentes e transformação dos dados.

A criação de novos atributos tem como objetivo dar realce às informações latentes contida nos dados. Essa atividade é geralmente um processo criativo que se baseia a partir de informações advindas da EDA. Na Tabela 5.3 do capítulo 5, estão descritos os atributos criados para o estudo de caso conduzido nesta tese.

Para o preenchimento dos dados faltosos, sugere-se analisar individualmente cada atributo. Os atributos com mais de 25% dos dados ausentes deverão ser removidos. Na Tabela 5.4 do Capítulo 5, é apresentada qual a heurística utilizada em cada um dos atributos para o estudo de caso conduzido nesta tese.

Após o preenchimento dos dados faltosos, os dados numéricos no domínio dos reais (como notas, frequência e renda bruta familiar) devem ser discretizados em uma escala de 0 a 10, a fim de diluir o fator ruído. Como isso, também se obtém os dados numa mesma escala.

É importante destacar que antes de escalonar os dados, é necessário verificar a existência de *outliers*. Caso existam *outliers* nos dados, eles devem ser removidos para não comprometerem as análises. Muitos dados de *outliers* ocorrem por falhas humanas em preenchimentos dos formulários dos dados.

Após a transformação inicial descrita acima, a qualidade dos dados é analisada novamente, a fim de remover atributos redundantes, com uma quantidade elevada de valores faltosos ou correlacionados.

O último passo da fase de transformação consiste na conversão dos dados categóricos para valores tipo binário.

4.4 Seleção de atributos

Uma das consequências da etapa anterior é o aumento da dimensionalidade dos dados, principalmente devido à transformação dos dados para o tipo binário. Portanto, esta etapa tem como principal objetivo selecionar os melhores atributos para minorar o problema do aumento da dimensionalidade dos dados.

É importante ressaltar que o processo de seleção dos atributos ocorre inicialmente na atividade de EDA, cujas análises indicam se há atributos correlacionados. Nesse segundo momento de seleção de atributos, propomos a utilização das seguintes técnicas:

- Matriz de Correlação: para remover todos os atributos que ainda estiverem com uma correlação elevada. Sugerimos o valor de 0,95 como limiar.
- Teste ANOVA para os dados numéricos: para remover dados que tenham pouca relação com a classe (p-value < 0,05, ou seja, rejeitar a H0. A Hipótese nula (H0) assume que as variâncias são iguais. Se houver variância igual entre atributo e a classe, significa que o atributo NÃO tem impacto na resposta. Se rejeitar a H0, significa que o atributo tem impacto na classe) pela métrica do ANOVA.
- CHI-QUADRADO para os dados categóricos (em formato binário): para remover aqueles que tenham pouca dependência com a classe (p-value < 0,05, ou seja, rejeitar a H0. A Hipótese nula assume que o atributo é INDEPENDENTE da classe. Se rejeitar a H0, significa que o atributo tem dependência com a classe) pela métrica do CHI-QUADRADO.
- VIF: para remover multicolinearidade a partir do teste VIF com valor acima de 10.

Destacamos que após o modelo de aprendizagem treinado (etapa modelos), é utilizada a técnica de RFEC sobre o melhor modelo obtido, a fim de verificar se há possibilidade de diminuir a quantidade de atributos sem perda considerável no desempenho do modelo.

4.5 Balanceamento de dados

Uma vez definida a base de dados, ela é dividida em 2 conjuntos estratificados: conjunto de treinamento e conjunto de testes. O primeiro conjunto de dados deverá ser utilizado para a seleção dos melhores hiper-parâmetros e treinamento dos modelos de aprendizagem. Uma vez que estamos diante de um problema de classes desbalanceadas, é necessário aplicar técnicas de balanceamento de dados sobre o conjunto de treinamento. Nesta tese, sugerimos a adoção das técnicas *Undersample* e SMOTE, uma vez que as abordagens baseadas em amostragem foram as mais utilizadas na Revisão Sistemática da Literatura (seção 3.1). Não há necessidade de se aplicar técnicas de balanceamento de dados sobre conjunto de testes, pois é importante mantê-los com a distribuição original, a fim de ter uma avaliação mais realística.

4.6 Modelo de aprendizagem

Cumprindo o critério de selecionar modelos mais interpretáveis, nessa tese não utilizamos técnicas de aprendizagem supervisionada do tipo caixa-preta, como as redes neurais, SVM, ou cinzas, como RF. Apesar dos modelos caixa-preta alcançarem bons resultados, eles não propiciam interpretações das relações entre os parâmetros de entrada. Assim, sugerimos o uso de árvores de decisão e regressão logística, que permitem a análise interpretativa dos relacionamentos das variáveis independentes com a variável alvo (alunos evadidos X alunos persistentes).

4.7 Avaliação

Para treinamento e otimização dos hiper-parâmetros dos modelos, sugerimos a utilização das métricas de avaliação *G-mean*, acurácia balanceada, MCC, precisão, *Recall*, F1, AUC, sendo a precisão, *Recall* e o F1 referentes aos alunos evadidos. Sugere-se calcular as métricas a partir da estratégia de seleção de dados de treinamento *cross-validation*, com o número de *folds K* igual a 10, como recomendado na literatura (Kohavi 1995). Para a busca dos melhores hiper-parâmetros dos modelos, sugere-se o emprego de busca exaustiva com métrica principal de avaliação o *G-mean*. Deve-se aplicar sobre o conjunto de testes a mesma métrica adotado no treinamento. Para comparação dos desempenho entre os modelos obtidos, sugerimos utilizar os testes 5X2 CV e o do McNeimar com significância de 0,05.

4.8 Interpretação

Para interpretação dos modelos gerados, sugere-se utilizar o gráfico em barra dos coeficientes da regressão logística e visualização da árvore de decisão gerada. A partir dessas visualizações, é possível destacar quais atributos são mais relevantes na tomada de decisão para cada uma das classes.

Visando realizar uma validação inicial do processo sistemático proposto nesta tese, no próximo capítulo será conduzido um estudo de caso tomando como base os dados escolares e socioeconômicos de estudantes de cursos integrados do Instituto Federal do Rio Grande do Norte (IFRN). Nesta seção, serão apresentados: a configuração do experimento e os resultados obtidos. Esse estudo de caso se mostra relevante na medida que possibilita se evidenciar os pontos fortes e fracos da presente proposta, viabilizando o seu aprimoramento e generalização, além de embasar escolhas de métodos e procedimentos a serem adotados para realizar as atividades integrantes do método proposto.

Capítulo 5

Estudo de caso

Com o propósito de avaliação do processo proposto e das técnicas associadas em cada uma das etapas, será conduzido um estudo de caso sobre evasão escolar a partir de dados que contêm atributos dinâmicos e estáticos de alunos do Instituto Federal do Rio Grande do Norte. O uso dos dados foi autorizados pela Pró-Reitora de Ensino do IFRN e, a partir dos dados brutos fornecidos, construímos a base que foi utilizada nesse trabalho. Essa instituição de educação é localizada no nordeste do Brasil e é distribuída por 20 *Campi* em diferentes cidades (Apodi, Caicó, Canguaretama, Ceará-Mirim, Currais Novos, Ipanguaçu, João Câmara, Lajes, Macau, Mossoró, Natal-Central, Natal-Cidade Alta, Natal-Zona Norte, Nova Cruz, Parelhas, Parnamirim, Pau dos Ferros, Santa Cruz, São Gonçalo, São Paulo do Potengi). Neste capítulo, serão descritos a base de dados utilizada, as ferramentas e os resultados alcançados em cada uma das etapas do processo proposto.

5.1 Base de dados

Após a fase de engenharia de atributos (ou seja, já realizadas algumas atividades de pré-processamento, como a transformação de atributos categóricos em valores binários), a base utilizada neste trabalho é composta por 7342 instâncias de alunos do ensino integrado (ensino médio com formação em educação profissional através de cursos técnicos com duração de quatro anos, na modalidade presencial) e por 137 atributos (excluindo o atributo classe). Os dados foram obtidos diretamente do sistema acadêmico oficial da instituição [1]. A última atualização dos dados utilizados foi em janeiro de 2018.

Na Tabela 5.1, estão descritos os 52 dos 137 atributos utilizados como variáveis independentes no modelo de aprendizagem, mais o atributo que representa a classe (atributo "evasao"). Esses 52 atributos são as variáveis escolhidas na fase de seleção de atributos (todas as outras variáveis independentes foram descartadas no treinamento do modelo de aprendizagem). Desses, 4 são dinâmicos relacionados ao histórico escolar do aluno, como a performance e a frequência; 48 são estáticos relacionados a informações demográficas, socioeconômicas, curso e o *campus* do aluno.

[1] https://suap.ifrn.edu.br/

Tabela 5.1: Descrição de atributos.

Atributo	**Descrição**
evasao	Tipo inteiro, sendo Valor "1"para alunos com status de "Evasão", "Cancelado", "Cancelamento Compulsório". Valor "0"para alunos com status de "Concluído", "Matriculado", "Matrícula Vínculo Institucional"
qnt_salarios	Tipo inteiro com valores de 1 a 10 representando a quantidade de salários mínimos (valor do ano de 2018 de 954 reais) derivado da renda bruta familiar
M_mean_int	Tipo inteiro com a média aritmética de todas as disciplinas no primeiro ano. Os valores estão discretizado entre 1 e 10.
M_std_int	Tipo inteiro com o desvio padrão, discretizado entre 1 e 10.
F_mean_int	Tipo inteiro com a frequência média do aluno, discretizado entre 1 e 10.
F_std_int	Tipo inteiro com o desvio padrão da frequência do aluno discretizado entre 1 e 10.
ano_nascimento	Tipo inteiro com ano de nascimento do aluno
ensino_fundamental_ conclusao	Tipo inteiro com o ano de conclusão do ensino fundamental
ano_letivo_ano	Tipo inteiro com o ano de ingresso do aluno
descricao_mae_ escolaridade	Tipo inteiro com a escolaridade da mãe
descricao_pai_ escolaridade	Tipo inteiro com escolaridade do pai
descricao_responsavel_ escolaridade	Tipo inteiro com escolaridade do responsável financeiro
qnt_pessoas_ domicilio	Tipo inteiro com quantidade de pessoas no mesmo domicílio
solteiro	Tipo verdadeiro ou falso (V/F) se o estado civil do aluno é solteiro
companhia_outros	V/F se o aluno mora com outras pessoas que não sejam familiares
companhia_parentes_ amigos	V/F se o aluno mora com parentes e amigos
companhia_ sozinho	V/F se o aluno mora só
companhia_pais	V/F se o aluno mora com os pais
sem_estudar_ NAN	V/F para valor não informado se o aluno parou de estudar
sem_estudar_sim	V/F se o aluno ficou sem estudar
res_nao_ informado	V/F para residência não informada
res_urbana	V/F se o aluno reside em zona urbana
statusImovelNo informado	V/F se o status do imóvel não foi informado
statusImovel Financiado	V/F se o aluno mora em imóvel financiado

statusImovel Pensionatoou Alojamento	V/F se o aluno mora em pensionato ou alojamento
trabAlunoNoinformado	V/F para trabalho do aluno não informado
trabAlunoNunca trabalhou	V/F se o aluno nunca trabalhou
respTrabNoinformado	V/F para trabalho do responsável financeiro não informado
respFinOprprioaluno	V/F se o aluno é o próprio responsável finaceiro
resFinPai	V/F se o responsável financeiro é o pai
respFinCnjuge	V/F se o responsável financeiro é o cônjuge
respFinTioa	V/F se o responsável financeiro é o tio ou tia
MC	V/F se o aluno estuda no campus Macau
JC	V/F se o aluno estuda no campus João Câmara
CN	V/F se o aluno estuda no campus Currais Novos
PAR	V/F se o aluno estuda no campus Parnamirim
NC	V/F se o aluno estuda no campus Nova Cruz
MO	V/F se o aluno estuda no campus Mossoró
CNAT	V/F se o aluno estuda no campus Natal Central
CM	V/F se o aluno estuda no campus Ceará Mirim
PF	V/F se o aluno estuda no campus Pau dos Ferros
SPP	V/F se o aluno estuda no campus de São Paulo do Potengi
TcnicodeNivelMdioemInformtica	V/F se o aluno cursa o técnico de Informática
TcnicodeNvelMdioemAlimentos	V/F se o aluno curso o técnico de Alimentos
TcnicodeNvelMdioemMeioAmbiente	V/F se o aluno cursa o técnico de Meio Ambiente
TcnicodeNvelMdioemQumica	V/F se o aluno cursa o técnico de Química
TcnicodeNvelMdioemMecatrnica	V/F se o aluno cursa o técnico de Mecatrônica
TcnicodeNvelMdioemTxtil	V/F se o aluno cursa o técnica em Textil
TcnicoemProgramaodeJogosDigitais	V/F se o aluno estuda o técnico de Jogos digitais
TcnicodeNvelMdioemControleAmbiental	V/F se o aluno cursa o técnico de Controle ambiental
TcnicodeNvelMdioemGeologia	V/F se o aluno cursa o técnico de Geologia
TcnicodeNvelMdioemManutenoeSuporteemInformtica	V/F se o aluno cursa o técnico de Manutenção e suporte

conhecimento_idiomas	V/F se o aluno tem conhecimento em outros idiomas além do português

Fonte: Elaborada pelo autor.

5.2 Ferramentas

Todas as atividades do processo proposto foram implementados utilizando o ambiente de desenvolvimento Colab do Google (Bisong 2019). Abaixo, estão listados os pacotes utilizados em cada uma das etapas do processo.

1. Entender os dados
   (a) Pandas, Numpy, *pandas_profiling*: manipulação, visualização de dimensionalidade, descrição do tipo, levantamento do balanceamento dos dados;
   (b) *Searborn* e *Matplot*: geração de gráficos estatísticos (Violino, Box-plot, Q-Q, Histogramas, análise de dependência, matriz correlação);
   (c) *Folium*: mapa geográficos;
   (d) *Prince*: mapa perceptual para análise de independência;
   (e) *Msno*: visualização de dados ausentes;
   (f) *Yellobrick*: visualização do T-SNE.
2. Engenharia de atributos
   (a) Pandas: preenchimento de dados ausentes, transformação de dados para valores binários;
3. Seleção de atributos
   (a) Pandas e *pandas_profiling*: seleção de atributos pela correlação;
   (b) *scikit-learn*: seleção de atributos por ANOVA, Chi-quadrado e RFECV;
   (c) *statsmodels*: seleção de atributos pelo VIF.
4. Balanceamento
   (a) *imbalanced-learn*: balanceamento de dados pelo SMOTE e *Undersample*;
5. Modelo
   (a) *scikit-learn*: aprendizado supervisionado ;
   (b) *GridSearchCV*: otimização de parâmetros;
6. Avaliação
   (a) *scikit-learn*: métricas de avaliação, otimização dos parâmetros
   (b) *mlxtend*: comparação dos modelos pelo 5X2CV, McNemar
7. Interpretação
   (a) *export_graphviz*: visualização da árvore de decisão;
   (b) *Matplot*: visualização dos coeficientes da regressão logística.

O código para criação da base está disponível em (Barros 2020*a*) referente aos dados estáticos e (Barros 2020*b*) aos dados dinâmicos. Para análise exploratória dos dados, o código está em (Barros 2020*c*). Por fim, em (Barros 2020*d*) estão os códigos de seleção, balanceamento, modelo, avaliação e interpretação.

Nas próximas, seções serão descritos os resultados encontrados na aplicação das técnicas em cada uma das etapas do processo.

5.3 Resultados e discussão

O modelo de Predição Escolar neste trabalho é caraterizado como um problema de classificação entre dois grupos desbalanceados de estudantes: (I) um com a tendência de persistir (majoritário), e (II) com tendência a evasão (minoritário). Nessa seção, serão apresentados os resultados gerados a partir da execução de cada uma das etapas do processo proposto. É importante lembrar que o processo busca pelos modelos mais simples e fácil entendimento humano, mas que seja superior ao modelo *Baseline*: "aluno com potencial de evasão é aquele que sua performance média E sua frequência média estão inferiores ao desempenho mínimo estabelecido pela organização didática". A etapa de "Entender o Problema"será suprimida, uma vez que o objetivo de definir o problema já foi contemplado em seções anteriores (Seção 3 e 4.1).

5.3.1 Entender os dados

Antes da criação da base descrita na Tabela 5.1, algumas análises exploratórias de dados foram realizadas. Para os dados numéricos, em destaque as médias finais das disciplinas, foram gerados os gráficos de histograma (Figura 5.1a), gráfico tipo violino (Figura 5.1b), gráfico *box-plot* (Figura 5.1c) e gráfico Q-Q (Figura 5.1d). O gráfico *box-plot* tenta sumarizar os dados da turma mostrando a média, desvio padrão e os quartis. Entretanto, há mais informações que ficam ocultas nesse gráfico, as quais podem ser observadas pelos gráficos violino e o histograma. Nesses gráficos, é verificada a presença de um grupo maior de alunos acima da média, mas há um trecho com uma cauda longa para esquerda com um pequeno pico na nota zero. Ou seja, os dados não estão seguindo a distribuição normal. Tal fato é confirmado pelo gráfico Q-Q, no qual percebe-se a descontinuidade na ponta esquerda (devido ao pico de notas zero que repercute na curtose) e a ponta direita em formato de arco (devido à assimetria dos dados).

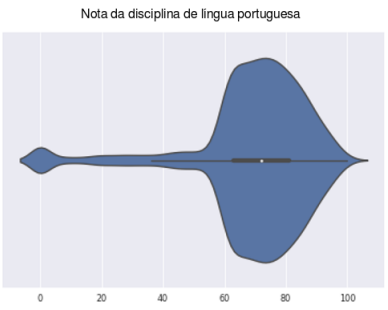
A distribuição conjunta dos dados das notas de português e matemática (Figura 5.2a), e o espalhamento das classes sobre essa distribuição (Figura 5.2b) também são analisados. Além disso, nos gráficos são verificados o aparecimento de dois grupos, de forma a mostrar que não há uma distribuição normal dos dados, além de explicitar que, apesar das notas se apresentarem com um bom indicador preditivo, elas sozinhas não são um fator determinante para evasão do aluno, visto que a classe que representa a evasão (círculos verdes) está espalhada ao longo das notas, inclusive com notas altas em ambas as disciplinas, mas de forma mais concentrada no grupo com notas abaixo da média.

A partir dos gráficos apresentados, é observada a presença de dois grupos. Um deles (com uma quantidade menor de instâncias) provavelmente relacionado aos alunos evadi-

Figura 5.1: EDA do atributo Nota da disciplina de Português de todos os alunos

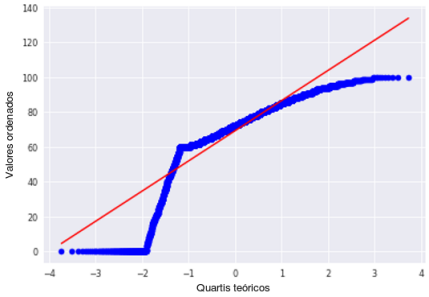
(a) Gráfico Histograma.

(b) Gráfico tipo violino.



(c) Gráfico *box-plot*.

(d) Gráfico Q-Q.



Fonte: Elaborada pelo autor.

dos e o outro (maior quantidade de instâncias) provavelmente com os alunos persistentes. Logo, surgem os primeiros indícios do desbalanceamento dos dados.

Como na análise anterior (Figuras 5.1 e 5.2), os dados categóricos foram analisados antes da transformação em atributos do tipo binário e enquadrados em uma única tabela. Para os atributos categóricos, é aplicada a técnica de análise de correspondência, visualizada a partir dos gráficos de mapa de calor e mapa perceptual, como nas Figuras 5.3 e 5.4. Os resultados apresentados na Tabela 5.2 são relacionados à classe dos alunos evadidos (nessa atividade, a Classe 0).

Tabela 5.2: Descrição de atributos para AC.

Atributo	**Descrição**
aluno_exclusivo_ rede_publica	Atração forte com o valor Verdadeiro
descricao_area_residencial	Atração forte com o valor “não informado”, e repulsão leve com "urbana"
descricao_companhia_ domiciliar	Atração forte com o valor “Cônjuge” e moderada com “Outros”
descricao_estado_civil	Atração forte com o valor “Divorciado”

descricao_historico	Atração forte com os cursos de “Informática” e “Têxtil” e moderada com “Meio Ambiente”
descricao_imovel	Atração forte com “Não informado” e de repulsão com "Financiado"
descricao_mae_ escolaridade	Atração forte com “Fundamental Incompleto”, moderado com “Alfabetizado”, "Não estudou", "Médio incompleto"e repulsão forte com "Médio Completo"
descricao_pai_ escolaridade	Atração forte com "Alfabetizado"e “Não estudou”, moderado com “Fundamental Incompleto” e repulsão forte com "Médio Completo"
descricao_raca	Atração forte com “Amarelo” e moderada com “Preta”, e de repulsão com "Indígena"
descricao_responsavel_ escolaridade	Atração forte com “Fundamental Incompleto” e “Alfabetizado”, e repulsão forte com "Médio Completo"e moderada com "Superior incompleto", "Pós graduação completo", "Pós graduação incompleto"
descricao_responsavel_ financeiro	Atração forte com “O próprio aluno” e moderada com “Cônjuge” e "Avô(ó)". Repulsão moderada com “Pai”
descricao_trabalho	Atração forte “Não informado”. Repulsão leve “Nunca trabalhou” (alunos que apenas estudam)
pessoa_fisica_sexo	Atração forte masculino. Repulsão forte feminino
possui necessidade especial	Atração forte “True”, repulsão moderada "False"
qtd_pessoas_domicilio	Atração acima de 6 e com o valor 0. Repulsão moderada com o valor 4
Sigla	Atração forte MC, moderada JC e leve CA, LAJ, NC, SC, SPP. Repulsão forte PAR, moderada para CN, e leve em CANG, CNAT, MO, PF, SGA
qnt_pc	Atração forte com 0, repulsão a partir de 1
qnt_salarios	Atração forte com renda bruta familiar de 1 salário mínimo. Repulsão a partir de 2.
tempo_entre_conclusao_ ingresso	Atração forte com 3 anos, repulsão com 1 ano (significa que não parou os estudos)

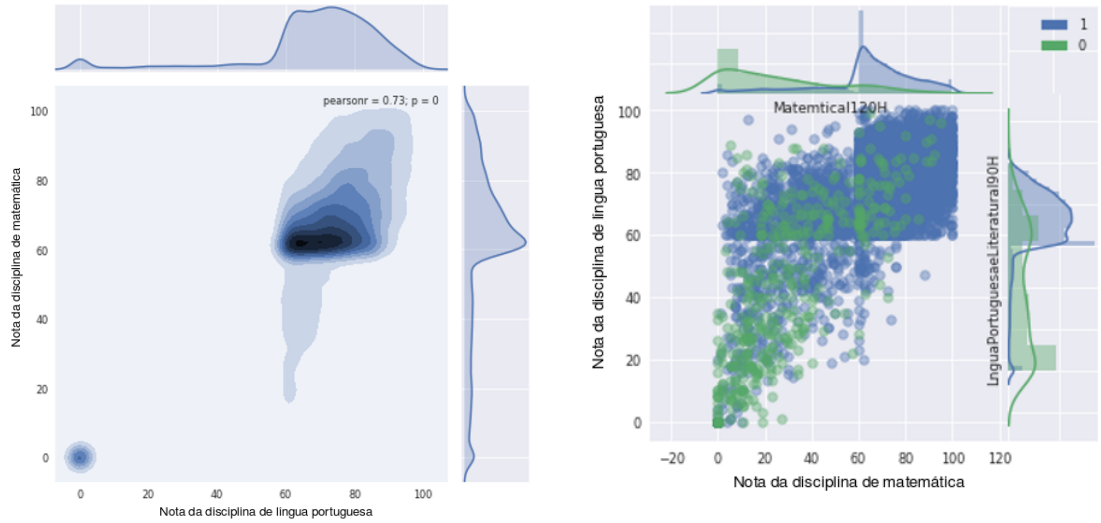
Fonte: Elaborada pelo autor.

A partir da análise da Tabela 5.2, são destacados as seguintes características do perfil do aluno evadido: alunos com baixo desempenho e pouca assiduidade, que já reprovaram ao menos uma vez, que não moram com os pais, com ensino fundamental em escola pública, os quais a escolaridade do responsável financeiro seja baixa, raça preta ou amarela, do sexo masculino, que convivam com mais de 6 pessoas no mesmo domicílio, que não tenha acesso a computador em casa, com renda bruta familiar de apenas 1 salário mínimo e que passaram mais de 2 anos entre a conclusão do ensino fundamental e o ingresso no ensino integrado do IFRN. Ou seja, há um forte indício que o problema da evasão do

Figura 5.2: Relação entre atributo Nota de Português e Matemática de todos os alunos

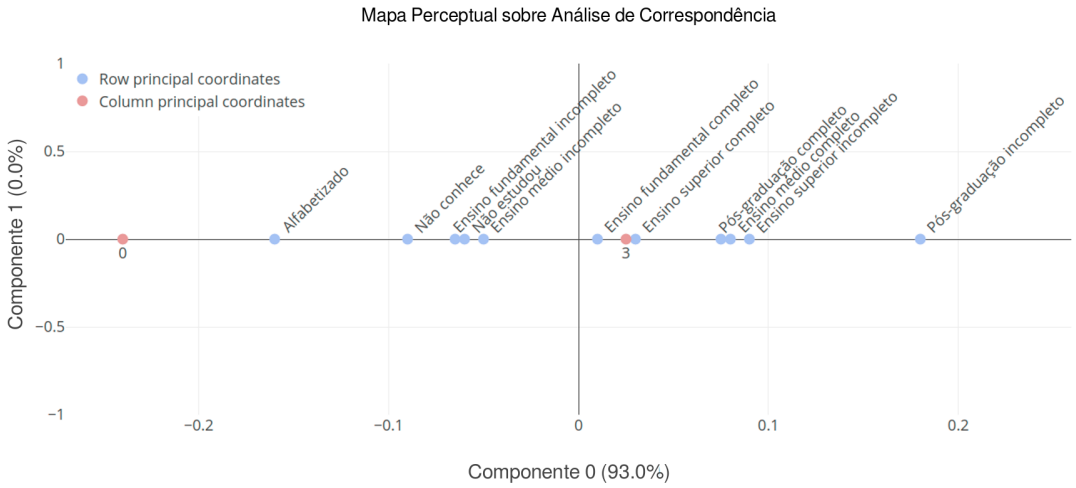
(a) Distribuição conjunta das notas português e matemática.

(b) Espalhamento das classes sobre distribuição conjunta.



Fonte: Elaborada pelo autor

Figura 5.3: Mapa perceptual do responsável financeiro.

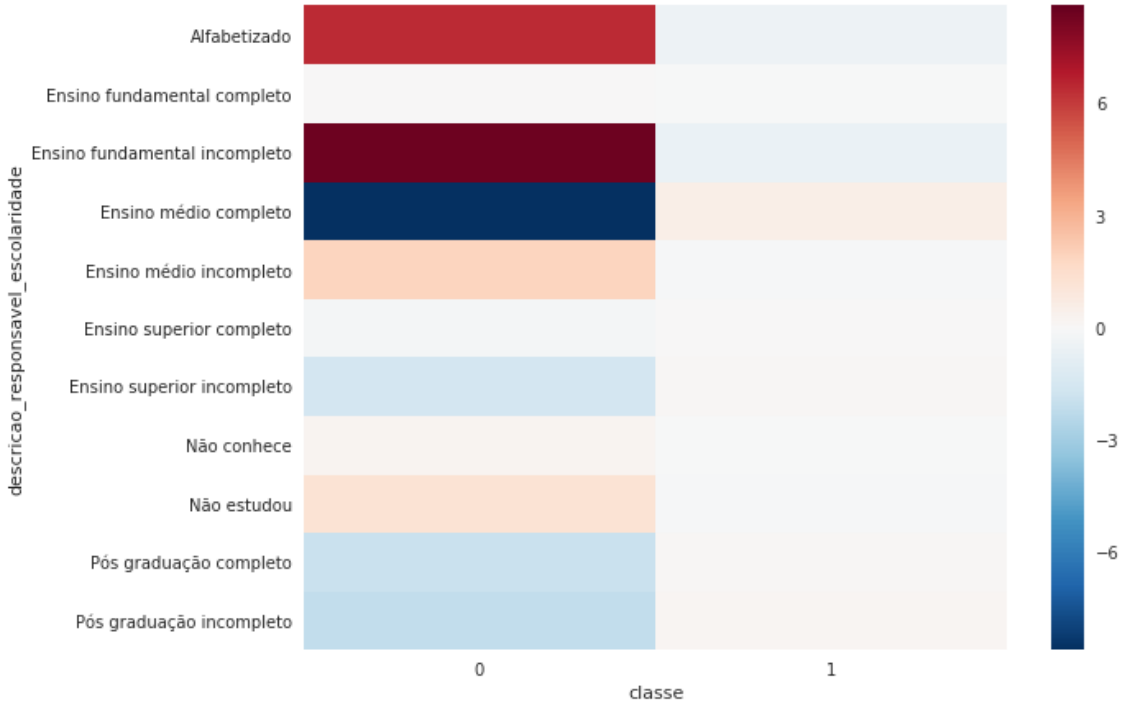


Fonte: Elaborada pelo autor.

integrado do IFRN esteja ligado à vulnerabilidade social do aluno.

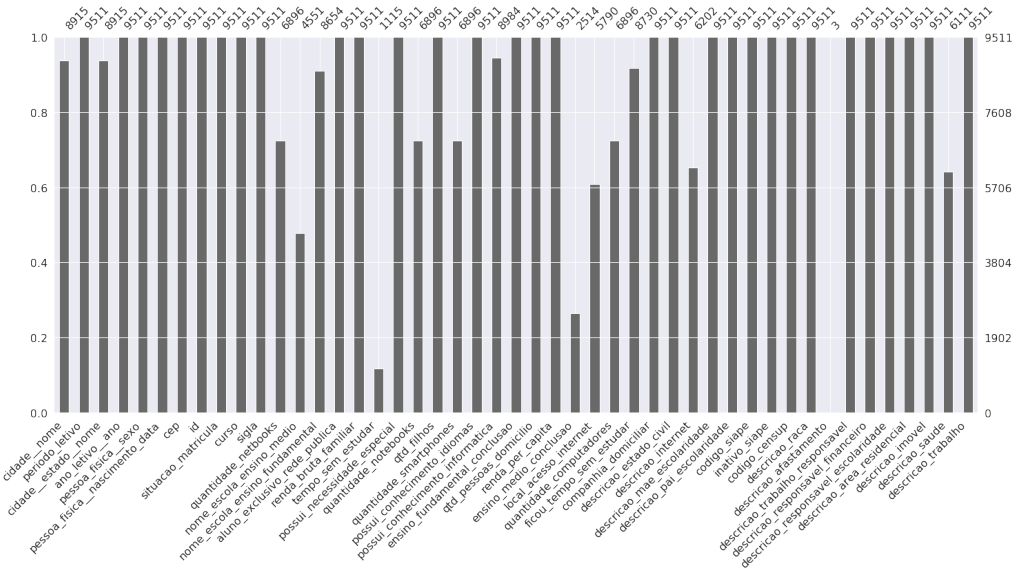
Após essa análise prévia, os dados são unificados em uma única tabela, para então realizar análises de dimensionalidade, os tipos de variáveis, ausência de dados (Figura 5.5), distribuição e estatísticas descritivas como média, mínimo, máximo.

Figura 5.4: Mapa de calor do responsável financeiro.



Fonte: Elaborada pelo autor.

Figura 5.5: Uso do *Msno* para análise dos dados ausentes

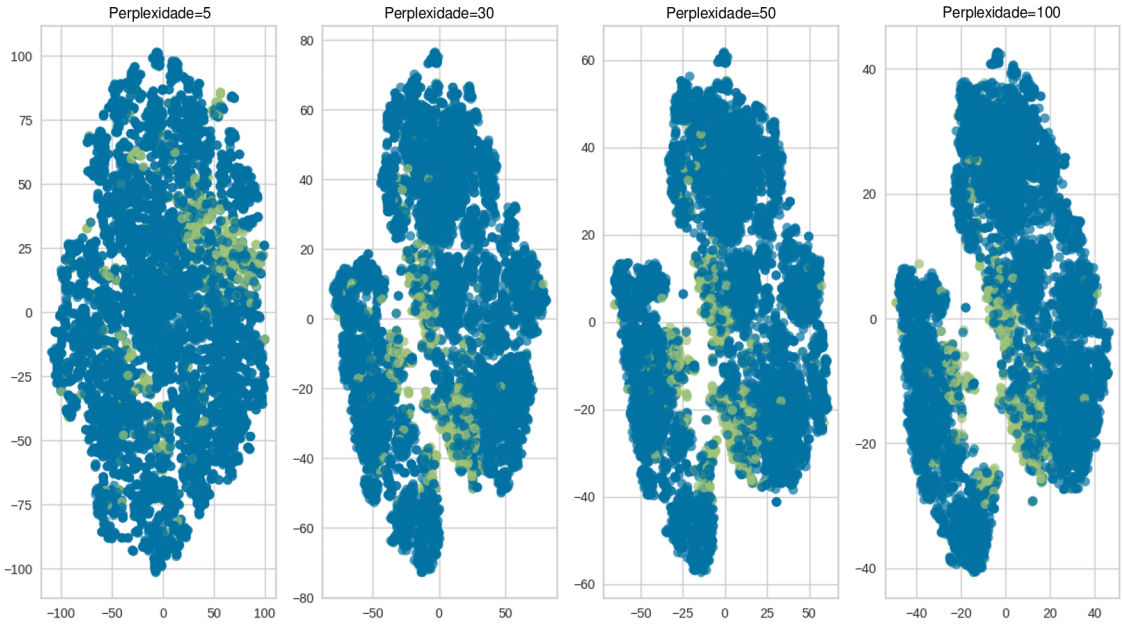


Fonte: Elaborada pelo autor

Essas ferramentas sugeriram a remoção de atributos com problemas de correlação ou valores constantes (16 atributos) e com ausência de dados acima de 25% (11 atributos), consistindo em excelentes indicadores para etapa de seleção de atributos.

Para uma visualização de toda a base de dados, foi utilizado o T-SNE, a fim de verificar se a visualização dos grupos de alunos evadidos e persistentes no plano bidimensional é possível. Na Figura 5.6, são utilizados vários valores de perplexidade (5, 30, 50, 100); nela, os alunos evadidos estão representados por bolas verdes, enquanto os persistentes, por bolas azuis. Apesar da separação dos dados em dois agrupamentos, visualmente os dados dos alunos evadidos estão espalhados pelos dois grupos.

Figura 5.6: Aplicação do T-SNE com diferentes perplexidades.

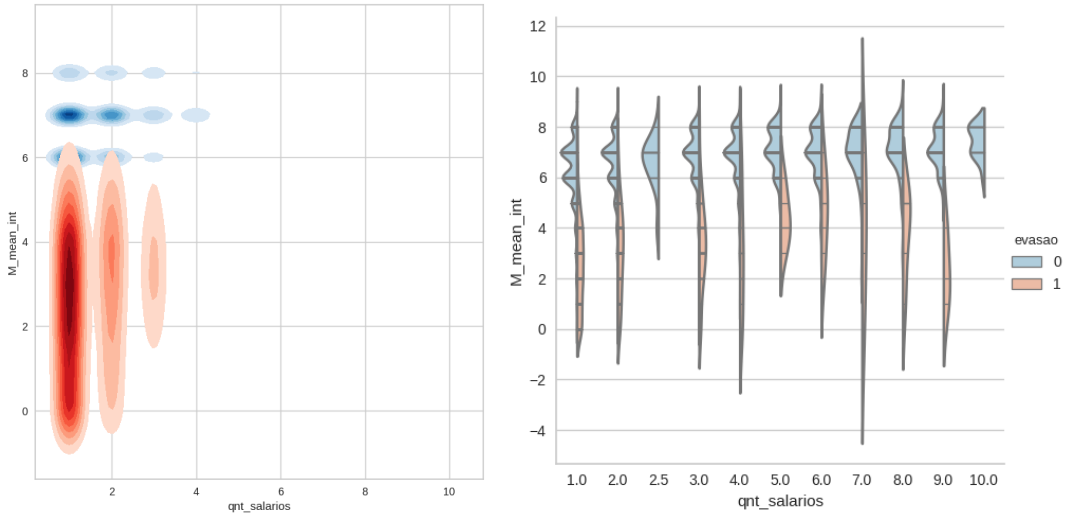


Fonte: Elaborada pelo autor.

Por fim, são feitos alguns gráficos para tentar encontrar padrões de interação entre as variáveis.

Na Figura 5.7, o eixo Y representa a média do aluno e o eixo X, a quantidade de salários mínimos de renda familiar. Ao longo do eixo Y, fica clara a divisão entre o grupo de alunos evadidos (em vermelho), com notas abaixo de 6, e grupo de alunos persistentes (em azul), com notas acima de 6. Entretanto, ao longo do eixo X, não parece haver grande diferença entre os dois grupos. O que se observa é que os exemplos vão minguando mais rapidamente no grupo de alunos evadidos, ao ponto que no último nível de renda (acima de 10 salários mínimos) praticamente só há alunos persistentes. Todavia, ao observar o Histograma da Figura 5.8 e relacionado a distribuição de dados entre os grupos de alunos evadidos e persistentes, apenas 7% dos alunos são evadidos, ou seja, essa diminuição mais rápida ao longo do eixo X pode estar relacionada à falta de instâncias desse grupo minoritário. Uma possível justificativa da não interferência da renda familiar na evasão é que

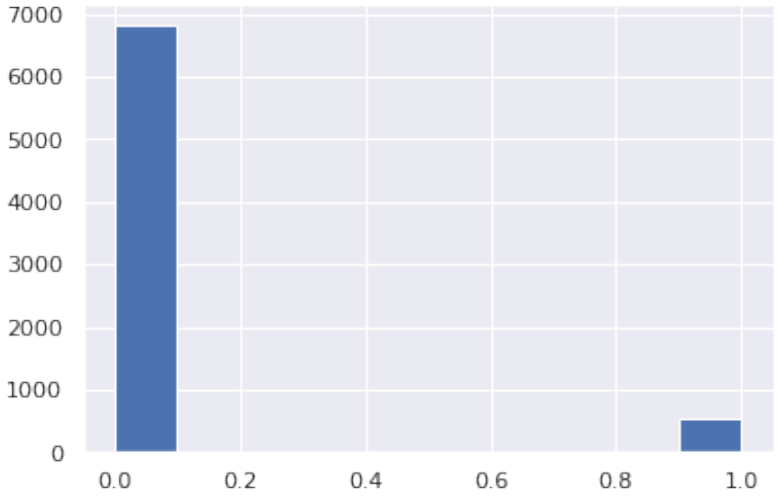
Figura 5.7: Visualização de interações entre variáveis.



Fonte: Elaborada pelo autor.

os alunos já foram selecionados no exame de seleção do IFRN. Ou seja, já são bons alunos (possivelmente os melhores da região) e as condições socioeconômicas influenciam menos, entretanto essa justificativa deve ser aprofundada em trabalhos futuros.

Figura 5.8: Histograma com a distribuição de alunos evadidos e persistentes.



Fonte: Elaborada pelo autor.

5.3.2 Engenharia de atributos

Após a fase de "Entender os dados", há uma melhor compreensão sobre a base disponível para então iniciar o processo de criação, preenchimento e transformação dos dados.

Uma vez que o cientista de dados tenha avaliado que as informações geradas até o momento não são suficientes para cumprir os critérios estabelecidos de conclusão, é necessário o tratamento da base de dados disponível. Na Tabela 5.3, estão descritos os atributos criados e sua descrição com o objetivo de tornar mais explícitos informações latentes nos dados.

Com a nova base de dados, são removidos todos os atributos com 25% de dados ausentes. Para os atributos com dados ausentes abaixo dessa porcentagem, foram utilizadas técnicas de preenchimento. Na Tabela 5.4, é apresentada qual a heurística utilizada em cada um dos atributos.

Após o preenchimento dos dados faltosos, os dados numéricos no domínio dos reais (Notas, Frequência e Renda Bruta Familiar) são discretizados em uma escala de 0 a 10, a fim de diluir o fator ruído. Como exemplo, a Figura 5.10 (a) mostra a distribuição de renda bruta familiar original a partir do gráfico tipo violino. No caso, há valores acima de 1 milhão de reais por mês, fato esse pouco provável de ocorrer, dado a realidade

Tabela 5.3: Novos atributos.

Atributo	**Descrição**
evasao	Tipo inteiro, sendo Valor "1"para alunos com status de "Evasão", "Cancelado", "Cancelamento Compulsório". Valor "0"para alunos com status de "Concluído", "Matriculado", "Matrícula Vínculo Institucional"
ano_nascimento	Tipo inteiro com ano de nascimento do aluno
qnt_salarios	Tipo inteiro com valores de 1 a 10 representando a quantidade de salários mínimos (valor do ano de 2018 de 954 reais) derivado da renda bruta familiar
M_mean_int	Tipo inteiro com a média aritmética de todas as disciplinas no primeiro ano. Os valores estão discretizado entre 1 e 10.
M_std_int	Tipo inteiro com o desvio padrão discretizado entre 1 e 10.
M_25_int	Tipo inteiro com primeiro quartil da média discretizado entre 1 e 10.
M_50_int	Tipo inteiro com segundo quartil da média discretizado entre 1 e 10.
M_75_int	Tipo inteiro com terceiro quartil da média discretizado entre 1 e 10.
F_mean_int	Tipo inteiro com a frequência média do aluno discretizado entre 1 e 10.
F_std_int	Tipo inteiro com o desvio padrão da frequência do aluno discretizado entre 1 e 10.

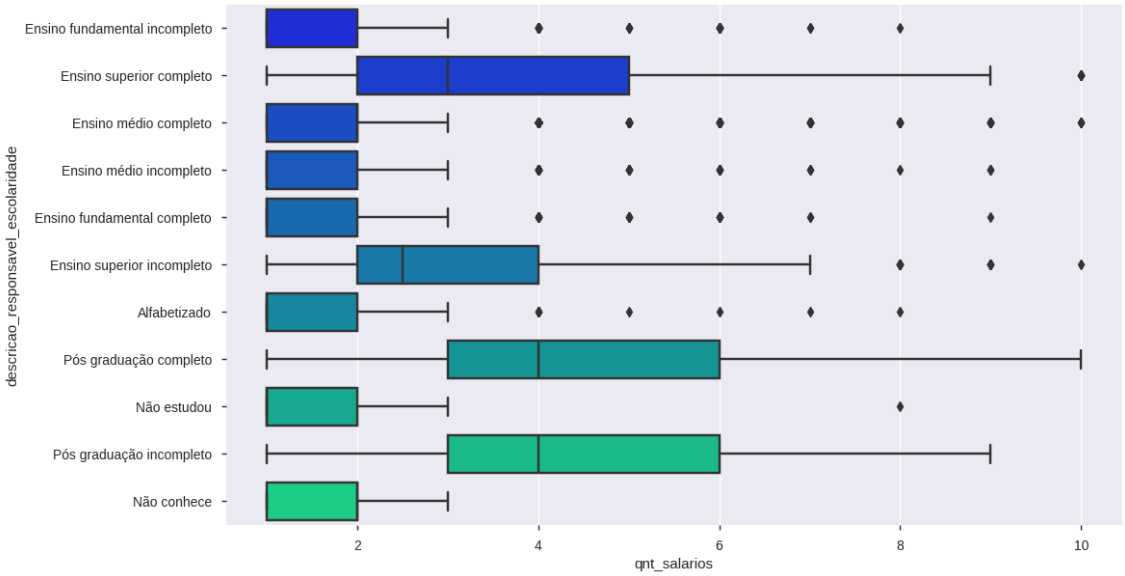
Fonte: Elaborada pelo autor.

Tabela 5.4: Preenchimento dos dados.

Atributo	**Descrição**
classe	removidas as instâncias sem a informação da classe
ensino_fundamental_ conclusao	preenchido com a média
qnt_salarios	média salarial de acordo com a escolaridade do responsável financeiro, como visto na Figura 5.9.
ficou_tempo_ sem_estudar	preenchimento com o valor "2"para representar o dado ausente.
possui_conhecimento_ informatica	preenchimento com o valor "2"para representar o dado ausente.
Média	todos as instâncias com dados ausentes em média foram removidas.
Frequência	toas as instâncias com dados ausentes na frequência foram removidas.

Fonte: Elaborada pelo autor.

Figura 5.9: Distribuição da renda bruta familiar em quantidade de salários por escolaridade do responsável financeiro.



Fonte: Elaborada pelo autor.

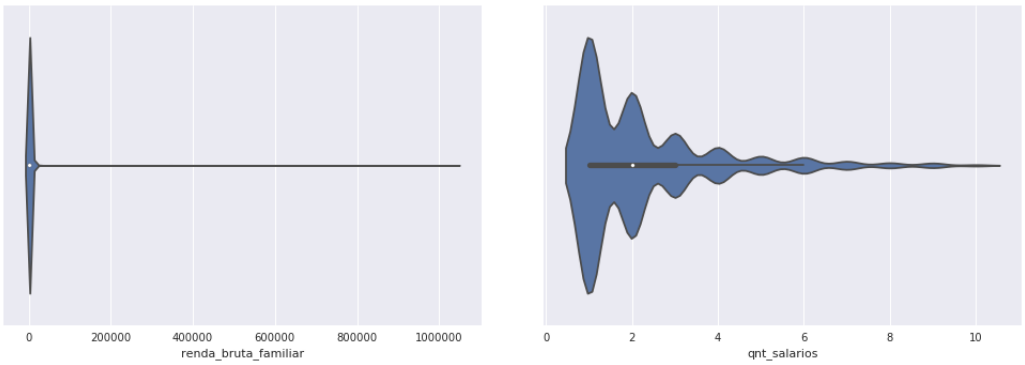
socioeconômica dos estudantes dessa base de dados. O resultado final após tratamento é mostrado na Figura 5.10 (b).

Por fim, os dados categóricos são então transformados em dados com valores binários. Os atributos convertidos são:

Figura 5.10: Renda familiar Bruta

(a) Distribuição da renda bruta familiar original dos estudantes.

(b) Renda familiar bruta após tratamento dos dados.



Fonte: Elaborada pelo autor

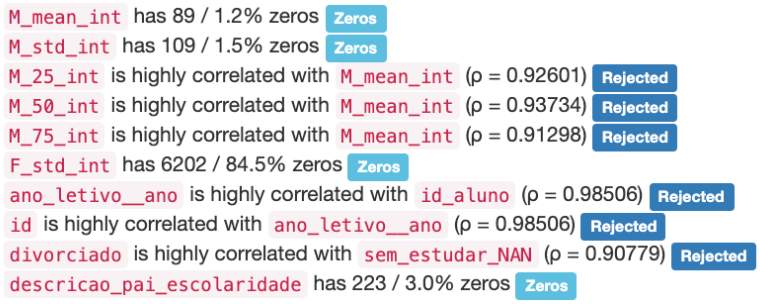
1. sexo do aluno;
2. *campus* do aluno;
3. ensino fundamental exclusivo em escola pública;
4. se o aluno ficou sem estudar;
5. se o aluno possui conhecimento em outra língua;
6. curso técnico do aluno;
7. se o aluno possui conhecimento de informática;
8. companhia domiciliar;
9. estado civil do aluno;
10. raça do aluno;
11. trabalho do responsável financeiro;
12. quem é o responsável financeiro;
13. área residencial (urbana, rural, quilombola, não informado);
14. tipo do imóvel (próprio, alojamento, financiado, emprestado, outro, não informado);

5.3.3 Seleção de atributos

Como já visto na seção 2.3.1, a etapa de engenharia de atributos causa o aumento da dimensionalidade da base, principalmente devido à transformação dos atributos categóricos para valores binários. Logo, a seleção de atributos é necessária, a fim de diminuir o ruído na base e facilitar a descoberta de padrões pelos modelos de aprendizagem. Todos os atributos com 25% de dados ausentes, com valores constantes e correlacionados indicados pelo *profile_pandas* foram removidos (Figura 5.11).

Os dados foram divididos em dois grupos: numéricos (12 atributos) e categóricos (125 atributos). Sobre os numéricos, é aplicada a técnica de ANOVA, a fim de verificar se há alguma correlação com a classe de evasão. Todos os atributos apresentaram correlação estatisticamente significativa. Sobre o grupo categórico, é aplicada a técnica de

Figura 5.11: Uso do *Profile_pandas* para visualização dos dados



Fonte: Elaborada pelo autor.

chi-quadrado e são selecionados aqueles que apresentavam dependência com a classe de evasão com significância de 0.05. Dos 125 atributos, foram selecionados 40.

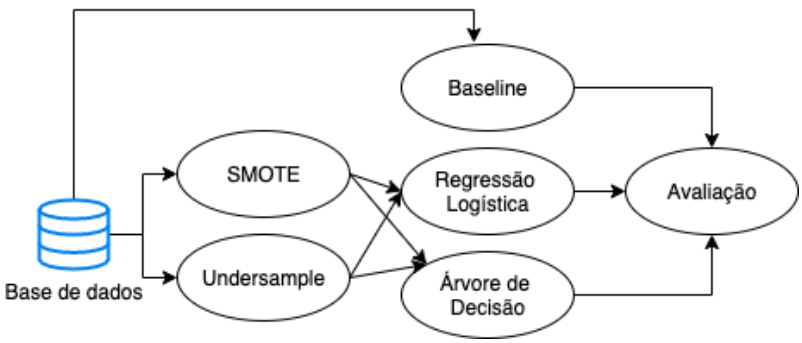
Por fim, foi aplicado o VIF em todos os atributos, a fim de verificar a multicolinearidade entre eles. Todos os atributos ficaram com valor abaixo de 6 e nenhum foi removido.

Todos os atributos selecionados estão descritos na Tabela 5.1.

5.3.4 Balanceamento de dados

A partir dos indícios encontrados na EDA sobre o desbalanceamento dos dados, foram aplicadas as técnicas de amostragem SMOTE e *Undersample*. Inicialmente, a base é dividida em 70% dos dados para treinamento e 30% dos dados para teste (por ser uma das proporções mais utilizadas, como visto na seção 2.5.1). Ambos os conjuntos foram estratificados de acordo com a proporção dos alunos evadidos. As técnicas de amostragem foram aplicadas sobre o conjunto de treinamento. Já o conjunto de testes se manteve desbalanceado, a fim de tornarem os valores das métricas de avaliação mais realísticos. A Figura 5.12 apresenta os 5 (cinco) experimentos realizados.

Figura 5.12: Fluxo de experimentos.



Fonte: Elaborada pelo autor.

Os experimentos são rotulados como:

1. LR-Under: Amostragem *UnderSample* e modelo de aprendizagem regressão logística.
2. DT-Under: Amostragem *UnderSample* e modelo de aprendizagem árvore de decisão.
3. LR-SMOTE: Amostragem SMOTE e modelo de aprendizagem regressão logística.
4. DT-SMOTE: Amostragem SMOTE e modelo de aprendizagem árvore de decisão.
5. BASELINE: Modelo *Baseline*.

5.3.5 Modelos de aprendizagem

Sobre o conjunto de treinamento, é realizada a seleção dos melhores parâmetros a partir da busca exaustiva, utilizando como métrica de avaliação o *G-mean* e a metodologia *cross-validation*, com K igual a 10.

Cada um desses modelos possui alguns parâmetros a serem otimizados. Na Tabela 5.5, estão descritos os parâmetros selecionados.

Tabela 5.5: Parâmetros dos modelos de aprendizagem.

Modelo:Parâmetro	**Descrição:Valores**
LR: solver	algoritmo de otimização: "newton-cg", "lbfgs", "liblinear"
LR e DT: class_weight	pesos associados às classes: None, "balanced"
DT: criterion	A função para medir a qualidade de uma divisão: "gini", "entropy"
DT: splitter	A estratégia usada para escolher a divisão em cada nó: "best", "random"
DT: max_features	O número de atributos a serem considerados ao procurar a melhor divisão: "auto", "sqrt", "log2"
DT: max_depth	A profundidade máxima da árvore: 1, 2, 3

Fonte: Elaborada pelo autor.

Na Tabela 5.6 são descritos os melhores parâmetros encontrados para cada um dos experimentos.

Para o modelo *Baseline* é definido que, para predição do aluno evadido, as duas condições seguintes devem ser cumpridas:

1. valor abaixo de 6 para o atributo M_mean_int E;
2. valor abaixo de 8 para o atributo F_mean_int.

5.3.6 Avaliação

Após o treinamento dos modelos de aprendizagem, é realizada sobre o conjunto de testes a avaliação a partir das métricas *G-mean*, acurácia balanceada, MCC, precisão, *Recall*, F1 e AUC. Também é utilizado o *cross-validation* estratificado com K igual a

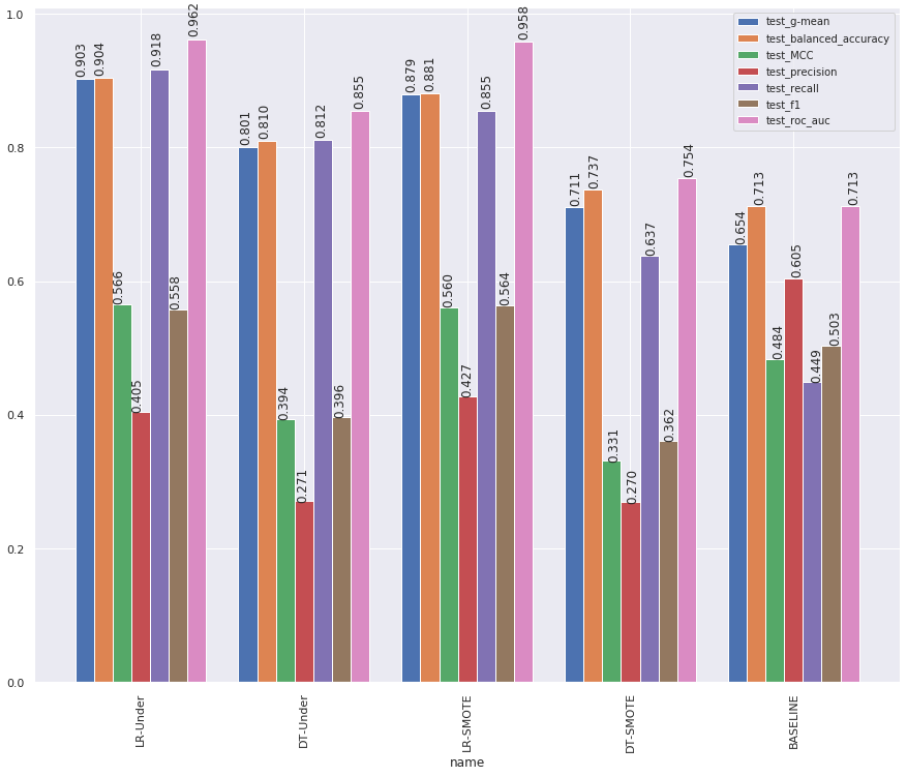
Tabela 5.6: Melhores parâmetros por modelo.

Modelo-Amostragem	**Parâmetros**	**G-mean**
LR-Under	'class_weight': None, 'solver': 'newton-cg'	0.898
DT-Under	'class_weight': None, 'criterion': 'gini', 'max_depth': 3, 'max_features': 'sqrt', 'splitter': 'best'	0.815
LR-SMOTE	'class_weight': None, 'solver': 'lbfgs'	0.901
DT-SMOTE	class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 3, 'max_features': 'log2', 'splitter': 'best'	0.805

Fonte: Elaborada pelo autor.

10 e as médias de cada uma das métricas, para cada experimento, estão apresentadas no gráficos 5.13 e na Tabela 5.7 (em azul estão marcados os maiores valores para cada métrica).

Figura 5.13: Média das métricas no conjunto teste.



Fonte: Elaborada pelo autor.

Tabela 5.7: Avaliação do modelo.

Métrica	MCC	AB	F1	G-mean	Precisão	Recall	AUC
LR-Under	0.566	0.904	0.558	0.903	0.405	0.918	0.962
DT-Under	0.394	0.810	0.396	0.801	0.271	0.812	0.855
LR-SMOTE	0.560	0.881	0.564	0.879	0.427	0.855	0.958
DT-SMOTE	0.331	0.737	0.362	0.711	0.270	0.637	0.754
BASELINE	0.484	0.713	0.503	0.654	0.605	0.449	0.713

Fonte: Elaborada pelo autor.

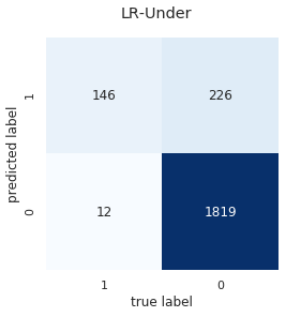
Ao observar as médias de cada um dos experimentos, das 5 (cinco) entre 7 (sete) métricas de avaliação, o LR-Under tem o maior valor de média (para o MCC, AB, *G-mean*, *Recall*, AUC). Dessas, o *G-mean*, a acurácia balanceada, o *Recall* e AUC ficam acima de 0.9. Todas as métricas mais robustas ao desbalanceamento (MCC, AB, *G-mean*, AUC) atingiram o seu maior valor nesse experimento. Entretanto, a métrica de precisão é inferior ao modelo *Baseline*. A precisão representa a qualidade da predição positiva (no trabalho os alunos evadidos), ou seja, a taxa de acerto de todas as instâncias previstas como positivas. Quando há uma diminuição nessa avaliação, significa que o erro de Falso Positivo (erro com a indicação que o aluno vai evadir, mas não evadiu) aumentou. Uma vez que foi definido neste trabalho ser mais importante identificar o aluno que evade frente ao aluno que persiste, o erro FP é menos importante que o Falso Negativo (erro que indica que o aluno vai persistir, mas evadiu). Logo, a degradação da precisão é um problema menor, já que as métricas robustas ao balanceamento do LR-Under foram as melhores entre os modelos. Outra consequência da precisão baixa é uma avaliação pior no F1 (uma vez que esta métrica é a média harmônica entre Recall e Precisão), apesar de a métrica *Recall* ser a melhor entre os modelos. A despeito de o LR-Under ter o melhor MCC, o valor não é alto (0.566). A justificativa é a mesma da precisão baixa: um erro de FP alto confirmado pela matriz de confusão (Figura 5.14a). A partir da matriz de confusão, também é possível identificar que o acerto das instâncias do aluno evadido foi a maior do LR-Under (146), o erro do FP (226) foi o segundo maior e o erro Falso Negativo (o mais importante nesse trabalho) foi o menor entre os modelos (12). Todas essas informações reverberam nas métricas como já explicado anteriormente.

O Modelo LR-SMOTE também obteve bons resultados, (terceira linha da Tabela 5.7), com avaliações superiores a todos os experimentos com a árvore de decisão, o que leva a crer que o modelo de regressão logística se adequa melhor a essa base de dados.

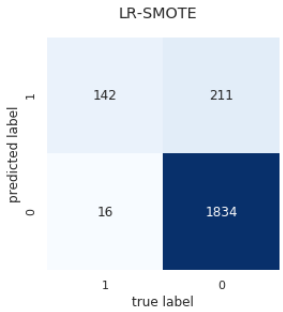
Os experimentos com árvore de decisão aparentam não ter sido superiores ao *Baseline*. Na Tabela 5.7, é possível verificar que o experimento DT-Under é inferior ao *Baseline* pelas métricas MCC, F1, Precisão. O experimento DT-SMOTE é inferior ao *Baseline* pelas métricas MCC, F1, precisão e apresenta valores muito próximos em AB e AUC. Ao analisar as matrizes de confusão (Figuras 5.14c e 5.14d), os maiores erros do FN (30 e 18 respectivamente DT-Under e DT-SMOTE) são maiores que os modelos de regressão logística, porém menores que o *Baseline* (87). O acerto dos alunos evadidos são maiores que o *Baseline* (128, 140, 71 respectivamente DT-Under, DT-SMOTE e

Figura 5.14: Matriz de confusão sobre o conjunto de treino

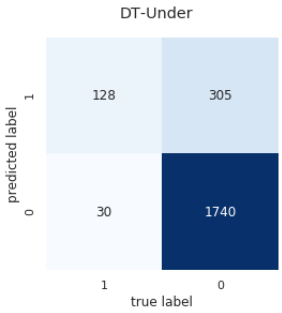
(a) LR-Under



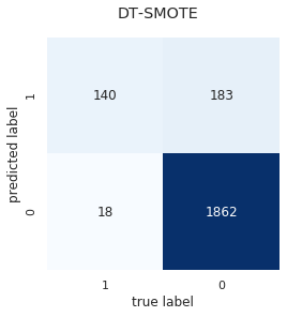
(b) LR-SMOTE



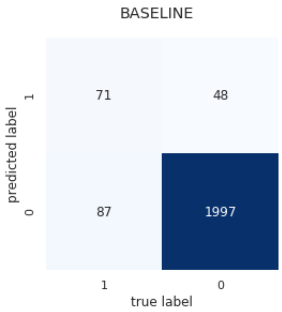
(c) DT-Under



(d) DT-SMOTE



(e) Baseline



Fonte: Elaborada pelo autor.

Baseline), entretanto menores que o da Regressão Logística.

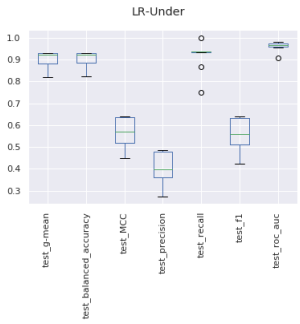
É importante lembrar que as árvores de decisão não são robustas em relação à variação dos dados, logo, elas tendem a variar sua estrutura de acordo com o conjunto de treinamento. É possível que o desempenho das árvores de decisão possa melhorar a partir do aumento de instâncias ao conjunto de treinamento.

Por fim, na Figura 5.15 são apresentados os valores das métricas de avaliação sobre o conjunto de TESTE em formato de *Box-Plot*. Isso é possível pois o *cross-validation* executa várias vezes (nesse trabalho, 10 para cada métrica de avaliação), permitindo gerar estatísticas descritivas. É interessante notar que os modelos de regressão logística (Figuras

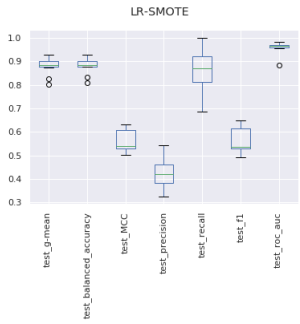
5.15a e 5.15b) apresentam uma menor variação do desvio padrão quando comparado aos outros modelos. Isso indica uma maior robustez sobre a variação das instâncias durante o *cross-validation*.

Figura 5.15: Box-plot das métricas de avaliação

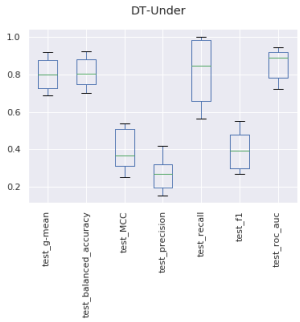
(a) LR-Under



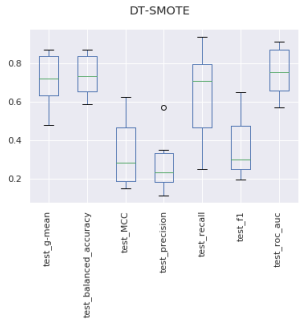
(b) LR-SMOTE



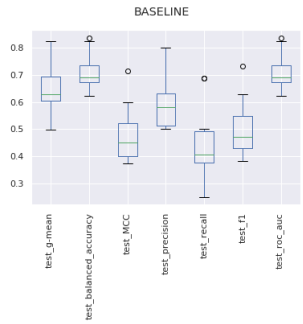
(c) DT-Under



(d) DT-SMOTE



(e) Baseline



Fonte: Elaborada pelo autor

Com o objetivo de verificar se há diferença estatística no desempenho entre os modelos de regressão logística e árvore de decisão contra o *Baseline*, são aplicados os testes 5x2CV e McNemar. A hipótese nula assume que os modelos possuem desempenhos iguais e a hipótese alternativa que os modelos possuem desempenhos diferentes. Caso o p-valor do teste tenha um valor menor o nível de significância (nesse trabalho de 0.05), a hipótese nula é rejeitada e assume a hipótese alternativa. Na Tabela 5.8, estão sumarizados

os valores dos testes. Na coluna "Interpretação"o valor "D"significa que os desempenho são diferentes e o valor "I"significa que os desempenhos são iguais. O primeiro "D"é referente ao teste McNemar e o segundo "D"é referente ao 5X2CV. Ambos os modelos de regressão logística indicam que os desempenho são diferentes quando comparados ao *Baseline*. Já para as árvores de decisão, o teste de 5x2CV (considerado o teste mais robusto) indica que não é possível rejeitar a hipótese nula, logo, não é possível afirmar que os desempenho são diferentes.

Tabela 5.8: Comparação entre modelos.

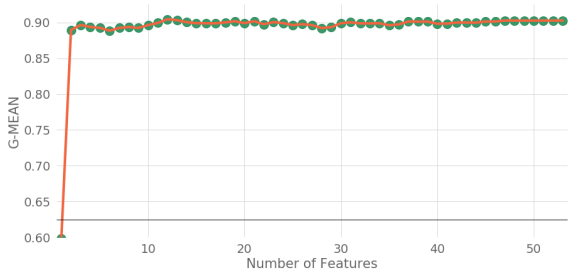
Modelo X *Baseline*	**McNemar: chi/p-value**	**5x2CV: stats/p-value**	**Interpretação**
LR-Under	90 / 1.404e-08	3.967 / 0.010	D,D
DT-Under	77 / 1.569e-08	0.770 / 0.476	D,I
LR-SMOTE	77 / 5.890e-06	3.506 / 0.017	D,D
DT-SMOTE	75 / 3.473-70	1.093 / 0.324	D,I

Fonte: Elaborada pelo autor.

A partir dos resultados apresentados, é possível afirmar que o modelo de regressão logística é superior ao *Baseline*, mas não é possível afirmar a mesma conclusão sobre o modelo de árvore de decisão.

Uma vez que o experimento de LR-UNDER teve um desempenho melhor do que os demais modelos, ele foi selecionado para uma otimização dos atributos através do RFECV, uma vez que o objetivo é ter o modelo mais simples e compreensivo possível. Após a execução da seleção de atributos pelo RFECV com métrica de avaliação *G-mean*, estratificado e K igual a 10, foram selecionados 12 atributos dos 52, como visto na Figura 5.16. Os atributos selecionados foram: 'M_mean_int', 'ano_nascimento', 'M_std_int', 'sem_estudar_NAN', 'res_nao_informado', 'JC', 'MO', 'CNAT', 'TcnicodeNvelMdioemTxtil', 'respFinCnjuge', 'TcnicodeNvelMdioemControleAmbiental', 'TcnicodeNvelMdioemGeologia'.

Figura 5.16: Curva da quantidade de parâmetros selecionados pelo RFEC.

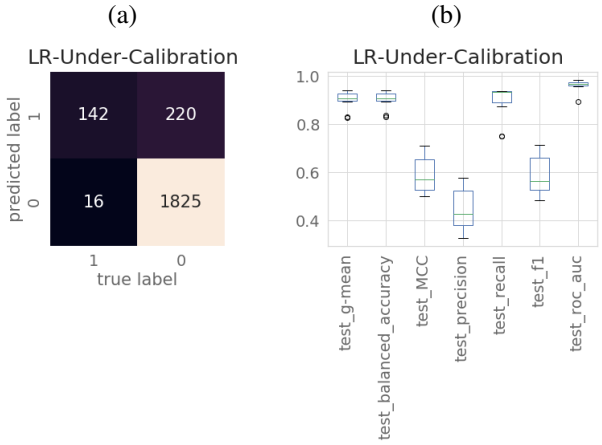


Fonte: Elaborada pelo autor.

Na Figura 5.17, é apresentado o desempenho do modelo após a redução das variáveis de entrada no conjunto de teste. Na Figura 5.18, é apresentada a média do *cross-*

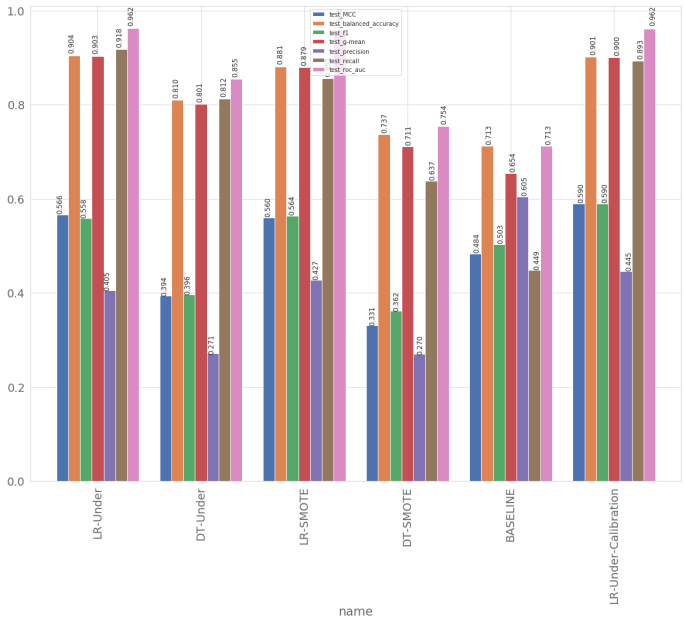
validation comparado aos demais modelos. Em todas as métricas mais robustas ao desbalanceamento (MCC, *G-mean*, Acurácia Balanceada e AUC), o modelo otimizado tem um desempenho melhor quando comparado ao *Baseline*.

Figura 5.17: Matriz de confusão e Box-Plot com as avaliações.



Fonte: Elaborada pelo autor.

Figura 5.18: Média das métricas no conjunto teste.



Fonte: Elaborada pelo autor.

Foi realizado o teste estatístico entre o Modelo Otimizado e o LR-Under, a fim de verificar se de fato os desempenhos são equivalentes. O teste realizado foi o McNemar,

pois ele é feito sobre a matriz de confusão de cada um dos modelos, independente dos atributos utilizados da base em cada modelo. O algoritmo utilizado para o Teste 5X2CV necessita que a base utilizada por cada modelo tenha os mesmo atributos, entretanto o modelo LR-Under utiliza 52 atributos e o modelo otimizado apenas 12, portanto, o teste 5X2CV não foi realizado. O McNemar resultou no valor chi-quadrado de 73 com p-valor de 0.628, ou seja, falhou na rejeição da hipótese nula, logo, não é possível afirmar que os desempenhos são diferentes.

A fim de trazer luz aos principais resultados dessa seção, abaixo segue uma sequência das atividades realizadas e os respectivos achados.

1. Foram feitos 5 experimentos (descritos na Seção 5.3.4), sendo 4 deles a combinação entre a técnica de aprendizado regressão logística e árvores de decisão, com as técnicas de balanceamento SMOTE e *Undersample*, além do modelo *Baseline*;
2. Os modelos inicialmente foram comparados a partir da média gerada pelo *cross-validation* das métricas *G-mean*, Acurácia Balanceada, MCC, Precisão, *Recall*, F1, AUC sobre o conjunto de testes (Figura 5.13 e Tabela 5.7) ;
3. O modelo com melhor desempenho pelas métricas mais robustas ao desbalanceamento (*G-mean*, Acurácia Balanceada, MCC, AUC) foi a regressão logística com *Undersample* rotulado como LR-Under;
4. Verificou-se que o modelo *Baseline* teve o desempenho superior pela métrica precisão frente ao LR-Under. Ao verificar a matriz de confusão (Figura 5.14), foi observado um aumento no erro Falso Positivo. Uma vez que foi definido neste trabalho ser mais importante identificar o aluno que evade frente ao aluno que persiste, o erro FP é menos importante que o Falso Negativo (erro que indica que o aluno vai persistir, mas evadiu). Portanto, não consideramos um problema grave o aumento do erro FP e a degradação da métrica precisão frente aos valores altos das métricas mais robustas ao desbalanceamento;
5. Os resultados da Regressão Logística com a técnica de balanceamento SMOTE foram superiores ao *baseline* e os modelos de árvore de decisão, entretanto não foram superior ao LR-Under. Uma vez que as DT são sensíveis aos dados de treinamento, para essa base, a regressão logística se apresenta como uma solução mais robusta, fato corroborado por um menor desvio padrão nas métricas (Figura 5.15), ou seja, menos sensível ao conjunto de treinamento;
6. A partir dos testes 5x2CV e McNemar, foi confirmado que há uma diferença estatística entre o modelo LR-Under e o *Baseline*. Portanto, é possível afirmar que o modelo LR-Under é superior ao *Baseline*.
7. Foi aplicada a técnica de seleção de atributos sobre o LR-Under, a fim de tornar o modelo mais simples. Dos 52 atributos, restaram apenas 12. O desempenho do modelo simplificado de 12 atributos de entrada é semelhante ao LR-Under.

5.3.7 Interpretação

Como já apresentado anteriormente, o objetivo de utilizar modelos de caixa-branca é permitir uma melhor interpretação pelo humano e extrair informações de como os atributos influenciam na predição. Para isso, nesse trabalho serão apresentados a visualização

das árvores de decisão geradas e os coeficientes da regressão logística do experimento LR-Under, o que teve o melhor desempenho entre os modelos. Apesar de a árvore de decisão não obter bons desempenhos quando comparada à regressão logística, é importante identificar quais são os atributos mais importantes para esse modelo e compará-los com a regressão, a fim de verificar se de fato há uma diferença que possa justificar um pior desempenho.

Na Tabela 5.9, são apresentados os atributos utilizados pelo experimento LR-Under, os valores de coeficientes, os valores de *odds* e a porcentagem efetiva. Eles estão ordenados pela influência sobre a classe de alunos evadidos, em que cada a cada 1 (uma) unidade no valor de entrada de um dado atributo influência para a predição do aluno como evadido, enquanto os outros atributos mantém os valores constantes. Por exemplo, o atributo "respFinPai"significa que o responsável financeiro pelo aluno é o pai, e caso esse valor seja "1", aumenta em 166.991% a chance em classificar o aluno como evadido. Já na outra ponta, o atributo "M_mean_int"representa que a cada 1 (uma) unidade que se acrescenta na nota, aumenta a chance em 76% do aluno ser classificado como persistente.

Tabela 5.9: Coeficientes do Modelo LR-Under.

Atributo	**Coeff**	**Odds**	**Porcentagem efetiva**
respFinPai	9.820e-01	2.670	166.991
sem_estudar_NAN	7.522e-01	2.122	112.168
TcnicodeNvelMdioemTxtil	7.484e-01	2.114	111.354
CNAT	6.140e-01	1.848	84.785
qnt_salarios	6.009e-01	1.824	82.374
respFinTioa	4.617e-01	1.587	58.679
NC	4.308e-01	1.539	53.853
trabAlunoNoinformado	3.280e-01	1.388	38.820
descricao_responsavel_escolaridade	3.219e-01	1.380	37.980
res_urbana	3.149e-01	1.370	37.011
TcnicoemProgramaodeJogosDigitais	2.738e-01	1.315	31.489
CM	2.617e-01	1.299	29.907
F_mean_int	2.074e-01	1.230	23.047
TcnicodeNvelMdioemInformtica	1.989e-01	1.220	22.011
TcnicodeNvelMdioemMeioAmbiente	1.950e-01	1.215	21.530
companhia_parentes_amigos	1.767e-01	1.193	19.330
respTrabNoinformado	1.741e-01	1.190	19.014
conhecimento_idiomas	1.125e-01	1.119	11.906
PF	5.867e-02	1.060	6.043
statusImovelNoinformado	4.038e-02	1.041	4.120
qtd_pessoas_domicilio	3.204e-02	1.033	3.256
JC	3.801e-03	1.004	0.381
statusImovelPensionatoouAlojamento	5.396e-05	1.000	0.005
TcnicodeNvelMdioemGeologia	0.000e+00	1.000	0.000
TcnicodeNvelMdioemQumica	0.000e+00	1.000	0.000

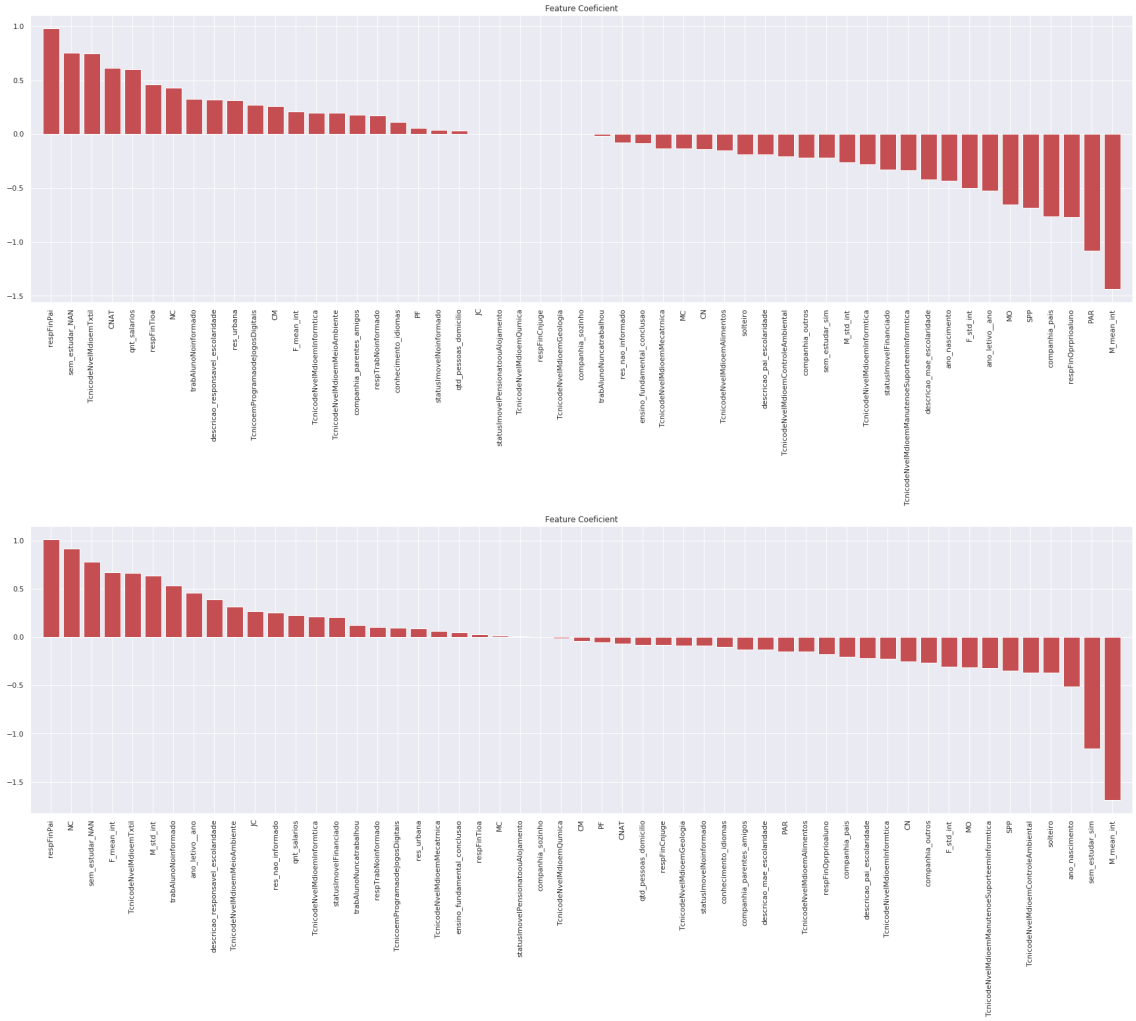
companhia_sozinho	0.000e+00	1.000	0.000
respFinCnjuge	0.000e+00	1.000	0.000
trabAlunoNuncatrabalhou	-1.621e-02	0.984	-1.608
res_nao_informado	-7.591e-02	0.927	-7.310
ensino_fundamental_conclusao	-8.321e-02	0.920	-7.984
TcnicodeNvelMdioemMecatrnica	-1.346e-01	0.874	-12.594
MC	-1.354e-01	0.873	-12.665
CN	-1.415e-01	0.868	-13.196
TcnicodeNvelMdioemAlimentos	-1.498e-01	0.861	-13.910
solteiro	-1.887e-01	0.828	-17.199
descricao_pai_escolaridade	-1.899e-01	0.827	-17.295
TcnicodeNvelMdioemControleAmbiental	-2.062e-01	0.814	-18.636
companhia_outros	-2.169e-01	0.805	-19.496
sem_estudar_sim	-2.171e-01	0.805	-19.517
M_std_int	-2.625e-01	0.769	-23.089
TcnicodeNivelMdioemInformtica	-2.779e-01	0.757	-24.266
statusImovelFinanciado	-3.302e-01	0.719	-28.121
TcnicodeNvelMdioemManutenoe Suporte-emInformtica	-3.366e-01	0.714	-28.579
descricao_mae_escolaridade	-4.202e-01	0.657	-34.311
ano_nascimento	-4.325e-01	0.649	-35.114
F_std_int	-4.990e-01	0.607	-39.288
ano_letivo__ano	-5.270e-01	0.590	-40.963
MO	-6.545e-01	0.520	-48.028
SPP	-6.804e-01	0.506	-49.360
companhia_pais	-7.630e-01	0.466	-53.372
respFinOprprioaluno	-7.718e-01	0.462	-53.781
PAR	-1.079e+00	0.340	-66.021
M_mean_int	-1.437e+00	0.238	-76.233

Fonte: Elaborada pelo autor.

É importante analisar que o processo proposto tem como objetivo gerar modelos preditivos, logo os coeficientes apresentados na Tabela 5.9 representam o quanto os atributos influenciam no resultado da predição. Em nenhum momento no trabalho, analisou-se se os atributos apresentados são a causa da evasão, logo, não é possível interpretar os valores dos coeficientes como o potencial do atributo em causar a evasão, mas sim que, para o modelo de regressão logística, o quanto o atributo contribui para classificar um aluno como evadido ou não. Na Figura 5.19, é apresentado o gráfico de barras dos coeficientes do experimento LR-Under (em cima) e LR-SMOTE (em baixo). Note que alguns atributos mudam de importância (por exemplo res_urbana ou sem_estudar_sim), mas há outros que se mantém importantes nos dois gráficos (por exemplo respFinPai e M_mean_int). Observe também que nos dois gráficos os atributos demográficos e socioeconômicos influenciam principalmente na predição para o aluno evadido, já o principal atributo que influencia para predição do aluno persistente é o atributo dinâmico que representa a mé-

dia acadêmica. Os 5 atributos que mais influenciam na predição do aluno evadido no experimento LR-Under são: "responsável financeiro pai", "não informou se ficou sem estudar", "faz o curso de têxtil", "estuda no CNAT"e "quantidade de salário". Já os 5 atributos que mais influenciam na predição do aluno persistente são: "média das disciplinas", "estuda em Parnamirim", "o responsável financeiro o próprio aluno", "mora com os pais"e "estuda em São Paulo do Potengi".

Figura 5.19: Coeficientes do LR-UNDER e LR-SMOTE.

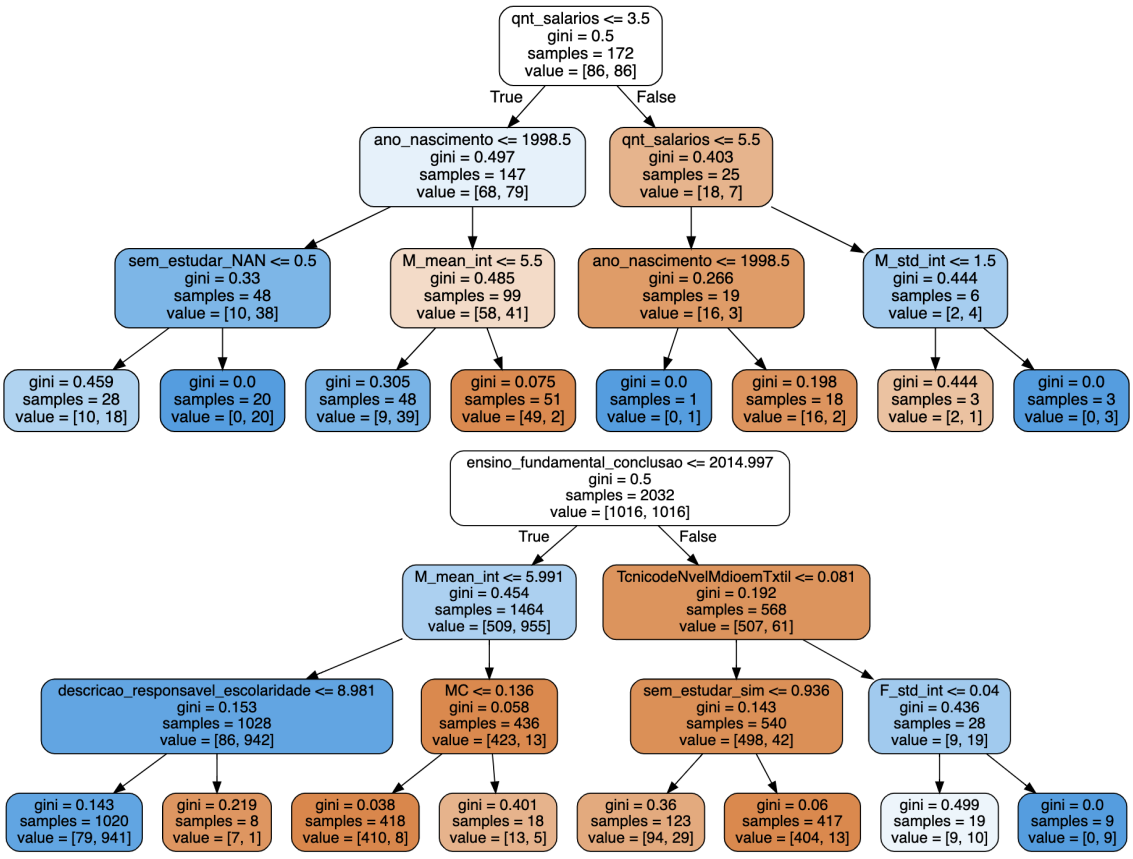


Fonte: Elaborada pelo autor.

Na Figura 5.20, são visualizadas as duas árvores de decisão (a primeira do experimento DT-Under e a segunda do DT-SMOTE). Quanto mais ao topo o atributo estiver localizado e a quantidade de vezes que o mesmo aparecer, mais importante o atributo é para o modelo. O atributo que representa a quantidade de salários é o mais importante para o DT-Under. Os outros atributos envolvidos são: o ano de nascimento, a média das notas, o desvio padrão das notas e que não informou se ficou sem estudar. Para o expe-

rimento DT-SMOTE, os atributos mais importantes foram: o ano de conclusão do ensino fundamental, a média das notas, se faz o curso de têxtil, a escolaridade do responsável, se estuda em Macau, ficou sem estudar e o desvio padrão da frequência escolar. Perceba que os atributos são bastante diferentes do experimento LR-Under, com exceção do atributo da média das notas, o curso de têxtil e que não informou se ficou sem estudar. Também é importante destacar que as árvores foram construídas sem o uso de todos os atributos disponíveis, ou seja, a exclusão de informações na construção comprometeu os desempenhos dos modelos.

Figura 5.20: Visualização das árvores dos experimentos DT-Under e DT-SMOTE.



Fonte: Elaborada pelo autor.

Para uma análise sobre como esses atributos de fato podem causar a evasão ou persistência, deve ser feito um estudo mais qualitativo, à luz de teorias educacionais.

Capítulo 6

Conclusão

Nessa tese, propusemos um processo sistemático baseado em técnicas de Ciência de Dados para geração de modelos de predição de evasão escolar. Para tal, foi definida uma sequência de etapas, a fim de modelar um fluxo de informação, que vai desde a definição do problema até a geração de informação útil a gestores e professores. Para cada etapa do processo proposto ("Entender o problema", "Entender os dados", "Engenharia de atributos", "Seleção de atributos", "Balanceamento de dados", "Modelos", "Avaliação e interpretação") foi realizada sua fundamentação teórica e, a partir do tipo dos dados, as técnicas e algoritmos mais adequados para o estudo de caso foram indicadas.

De forma sucinta, realizamos as seguintes atividades para cada etapa (sistematizada na Figura 4.1):

1. **Entender o problema**: foi realizada uma revisão sistemática da literatura, além da condução de entrevistas não-estruturadas com professores e gestores para definir o problema;
2. **Entender os dados**: foi realizada a Análise Exploratória de Dados, principalmente a partir da visualização de gráficos como: histograma, box-plot, violino, Q-Q, T-SNE, análise de correspondência e distribuição de dados.
3. **Engenharia de atributos**: foram criados novos atributos, preenchimento de valores ausentes, discretização de valores contínuos e transformação de valores categóricos em binários;
4. **Seleção de atributos**: foram removidos atributos numéricos correlacionados, atributos categóricos com baixa dependência com a classe e, por fim, com colinearidade;
5. **Balanceamento de Dados**: uma vez que foi verificado o desbalanceamento entre as classes, foram aplicadas sobre o conjunto de treinamento as técnicas de amostragem *Unsersampling* e SMOTE;
6. **Modelos**: a fim de cumprir com o princípio de selecionar modelos de mais fácil compreensão aos humanos, a regressão logística e as árvores de decisão foram utilizados como algoritmos de aprendizagem ;
7. **Avaliação**: os modelos foram avaliados a partir das métricas G-mean, AB, MCC, Precisão, *Recall*, F1 e AUC. Para comparação com o modelo *Baseline*, foram utilizados os testes McNemar e 5X2CV.
8. **Interpretação**: por fim, foram analisados os coeficientes da regressão logística e os nós das árvores de decisão treinadas, a fim de extrair conhecimento útil para

gestores e professores.

Como resultado, verificamos que o modelo com melhor desempenho nas métricas mais robustas ao desbalanceamento (MCC, AB, *G-mean*, AUC) foi a regressão logística, utilizando a técnica de balanceamento de dados *Undersample* (LR-Under). O seu desempenho superior foi validado pelos testes estatísticos 5X2CV e McNemar, quando comparado à hipótese: "o modelo preditivo de evasão escolar gerado tem um desempenho melhor frente a suposição que o aluno irá evadir caso: (I) sua performance média E (II) sua frequência média estejam inferiores ao desempenho mínimo estabelecido pela organização didática". É importante enfatizar que a regressão logística conseguiu trazer informação útil para gestores e professores, uma vez que ela é um modelo de caixa-branca, em que é possível verificar os parâmetros mais importantes para a tomada de decisão e o quanto ela contribui para isso.

Portanto, a proposta de aplicar um processo sistemático para geração de modelos de predição de evasão frente a um modelo preditivo generalista permitiu a criação de um preditor de evasão escolar adaptado aos dados disponíveis, com desempenho superior ao *Baseline* definido, além da possibilidade de interpretação e extração de conhecimento útil para gestores e professores. Também é importante destacar que, a partir da análise exploratória dos dados, nesse trabalho evidencia que o fenômeno de evasão escolar deve ser abordado como um problema de classes desbalanceadas, o qual deve utilizar-se de ferramentas e métricas apropriadas, a fim de gerar um modelo de predição robusto ao Paradoxo da Acurácia.

Entretanto, é importante enfatizar que o modelo *Baseline* teve o desempenho superior pela métrica precisão frente ao LR-Under. O motivo foi o aumento no erro Falso Positivo. Uma vez que foi definido neste trabalho ser mais importante identificar o aluno que evade frente ao aluno que persiste, o erro FP é menos importante que o Falso Negativo (erro que indica que o aluno vai persistir, mas evadiu). Portanto, não consideramos o aumento do erro FP e a degradação da métrica Precisão frente aos valores altos das métricas mais robustas ao desbalanceamento como um problema grave. Todavia, é uma limitação observada pela pesquisa, e um desafio futuro melhorar o desempenho sem aumentar o erro FP.

As contribuições da respectiva tese são listadas a seguir:

- Propor o processo para gerar um modelo preditivo de evasão escolar baseado em Ciência de Dados robusto ao problema de classes desbalanceadas e de fácil interpretação;
- Detalhar quais técnicas, algoritmos e modelos devem ser utilizados em cada etapa do processo;
- Demonstrar que a evasão escolar deve ser abordada como um problema de classes desbalanceadas, para evitar o fenômeno do Paradoxo da Acurácia;
- Realizar a Análise Exploratória de Dados dos alunos do integrado do IFRN e iniciar a discussão do perfil de alunos mais propensos a persistir e evadir;
- Construção de uma base de dados com informações reais.

Em relação a trabalhos futuros, planejamos aplicar o processo proposto em outras bases de dados educacionais, a fim de verificar se os modelos gerados conseguem se ade-

quar aos dados disponíveis com desempenho superior ao modelo *Baseline*. Além disso, pretendemos utilizar outras técnicas avançadas de aprendizado de máquina, como *Deep Learning* e programação probabilística, a fim de comparar a relação entre acurácia e interpretabilidade. Por fim, foram levantados algumas hipóteses que pretendemos investigar, também utilizando métodos qualitativos, sendo elas:

1. Verificar a relação de dependência alta entre vulnerabilidade social e evasão para o aluno do integrado do IFRN;
2. Verificar o quanto a renda bruta interfere ou não na evasão para o aluno do integrado do IFRN;

Referências Bibliográficas

Al-Jallad, N.T., X. Ning & M.A. Khairalla (2019), 'An interpretable predictive framework for students' withdrawal problem using multiple classifiers', *Engineering Letters* **27**(1), 1–8.
URL: *https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062952524partnerID=40md5=29b801f7af0da6966fa0476dd0f5a443*

Asif, Raheela, Agathe Merceron, Syed Abbas Ali & Najmi Ghani Haider (2017), 'Analyzing undergraduate students' performance using educational data mining', *Computers & Education* **113**, 177 – 194.
URL: *http://www.sciencedirect.com/science/article/pii/S0360131517301124*

Astin, Alexander w. (1999), 'Student involvement: A developmental theory for higher education', *Journal of College Student Development* pp. 518–529.

Barros, R. P (2017), Políticas públicas para a redução do abandono e da evasão escolar de jovens, Relatório técnico, Fundação Brava, Insper, Instituto Unibanco e Instituto Ayrton Senna.
URL: *http://gesta.org.br/wp-content/uploads/2017/09/Politicas-Publicas-para-reducao-do-abandono-e-evasao-escolar-de-jovens.pdf*

Barros, Thiago M., Ivanovitch Silva & Luiz Affonso Guedes (2017), 'Modelagem e visualização científica de dados educacionais: Estudo de caso sobre o desempenho em componentes curriculares', *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* **6**(1), 654.
URL: *https://www.br-ie.org/pub/index.php/wcbie/article/view/7451*

Barros, Thiago M., Ivanovitch Silva & Luiz Affonso Guedes (2018), 'Uso da técnica de análise de correspondência para análise exploratória de dados no contexto educacional', *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* **7**(1), 389.
URL: *https://br-ie.org/pub/index.php/wcbie/article/view/8264*

Barros, Thiago M., Ivanovitch Silva & Luiz Affonso Guedes (2019), 'Determination of dropout student profile based on correspondence analysis technique', *IEEE Latin America Transactions* **17**(09), 1517–1523.

Barros, Thiago M., Plácido A. Souza Neto, Ivanovitch Silva & Luiz Affonso Guedes (2019), 'Predictive models for imbalanced data: A school dropout perspective', *Education Sciences* **9**(4), 275.
URL: *http://dx.doi.org/10.3390/educsci9040275*

Barros, Thiago Medeiros (2018), 'Modelo ifrn integrado', https://github.com/tmedeirosb/modelo_ifrn_integrado/tree/master/versao_2.

Barros, Thiago Medeiros (2019), 'Modelo ifrn integrado'.
URL: *https://github.com/tmedeirosb/modelo_ifrn¡ntegrado/blob/master/PAPER_EXPERT_FINAL.ip*

Barros, Thiago Medeiros (2020*a*), 'Data wragling/dados estáticos', https://colab.research.google.com/drive/1DtBuP_uwX3ymp7iF1vq1WqWdBKK9kdns.

Barros, Thiago Medeiros (2020*b*), 'Data wragling/dados estáticos', https://colab.research.google.com/drive/1uwMFQi65-jxPSGwihDNcvCuYfQTRICFZ.

Barros, Thiago Medeiros (2020*c*), 'Data wragling/dados estáticos', https://colab.research.google.com/drive/1wM8F8qZRLQOcfrwjfGe_kNEm1VpKzl3X.

Barros, Thiago Medeiros (2020*d*), 'Data wragling/dados estáticos', https://colab.research.google.com/drive/1sRp8yYrZpsczEI_-j67TljDIPMQpkVS0.

Bean, John P. (1982), 'Conceptual models of student attrition: How theory can help the institutional researcher', *New Directions for Institutional Research* .

Bisong, Ekaba (2019), *Google Colaboratory*, Apress, Berkeley, CA, pp. 59–64.

Bolón-Canedo, Verónica, Noelia Sánchez-Maroño & Amparo Alonso-Betanzos (2016), 'Feature selection for high-dimensional data', *Progress in Artificial Intelligence* **5**(2), 65–75.
URL: *https://doi.org/10.1007/s13748-015-0080-y*

Brackett, Marc A., James L. Floman, Claire Ashton-James, Lillia Cherkasskiy & Peter Salovey (2013), 'The influence of teacher emotion on grading practices: a preliminary look at the evaluation of student writing', *Teachers and Teaching* **19**(6), 634–646.
URL: *https://doi.org/10.1080/13540602.2013.827453*

Bradley, Andrew P. (1997), 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern Recognition* **30**(7), 1145 – 1159.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320396001422*

BRASIL (2016*a*), 'Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro', http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior.
URL: *http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior*

BRASIL (2016*b*), Panorama da educação destaques do education at a glance 2016, Relatório técnico, DEED/MEC.
URL: *http://download.inep.gov.br/acoes_internacionais/eag/documentos/2016/panorama_da_educacao_2016_eag.PDF*

BRASIL (2018), 'Mec libera 100% do orçamento de custeio para universidades e institutos federais'.
URL: *http://redefederal.mec.gov.br/links/1204-mec-libera-100-do-orcamento-de-custeio-para-universidades-e-institutos-federais*

Brownle, J. (2019), 'How to choose a feature selection method for machine learning', https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/. Accessed: 2020-01-22.

Burgos, Concepción, María L. Campanario, David de la Peña, Juan A. Lara, David Lizcano & María A. Martínez (2018), 'Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout', *Computers Electrical Engineering* **66**, 541 – 556.
URL: *http://www.sciencedirect.com/science/article/pii/S0045790617305220*

Cady, Field (2017), *The Data Science Handbook*, 1thª edição, Wiley.

Cano, A. & J. D. Leonard (2019), 'Interpretable multiview early warning system adapted to underrepresented student populations', *IEEE Transactions on Learning Technologies* **12**(2), 198–211.

Cerqueira, D. & R. Moura (2019), 'Oportunidades laborais, educacionais e homicÍdios no brasil', *Instituto de Pesquisa Econômica Aplicada - IPEA* .
URL: *http://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td*$_2$514.*pdf*

Chakure, Afroz (2019), 'Introduction to machine learning', https://medium.com/swlh/introduction-to-machine-learning-different-types-of-machine-learning-alg Accessed: 2020-01-22.

Chen, Chao, Andy Liaw & Leo Breiman (2004), Using random forest to learn imbalanced data, Relatório técnico, Berkeley.
URL: *https://statistics.berkeley.edu/tech-reports/666*

Chicco, Davide & Giuseppe Jurman (2020), 'The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation', *BMC genomics* **21**(1), 6.

Corbett-Davies, Sam & Sharad Goel (2018), 'The measure and mismeasure of fairness: A critical review of fair machine learning', *CoRR* **abs/1808.00023**.
URL: *http://arxiv.org/abs/1808.00023*

Delen, Dursun (2011), 'Predicting student attrition with data mining methods', *Journal of College Student Retention: Research, Theory & Practice* **13**(1), 17–35.
URL: *https://doi.org/10.2190/CS.13.1.b*

Delen, Dursun, Kazim Topuz & Enes Eryarsoy (2019), 'Development of a bayesian belief network-based dss for predicting and understanding freshmen student attrition', *European Journal of Operational Research* .
URL: *http://www.sciencedirect.com/science/article/pii/S0377221719302954*

Demšar, Janez (2006), 'Statistical comparisons of classifiers over multiple data sets', *J. Mach. Learn. Res.* **7**, 1–30.

Dewey, John (1938), *Logic - The theory of inquiry*, HENRY HOLT AND COMPANY.

Dietterich, Thomas G. (1998), 'Approximate statistical tests for comparing supervised classification learning algorithms', *Neural Comput.* **10**(7), 1895–1923.
URL: *https://doi.org/10.1162/089976698300017197*

Domingos, Pedro (2012), 'A few useful things to know about machine learning', *Commun. ACM* **55**(10), 78–87.
URL: *https://doi.org/10.1145/2347736.2347755*

Enders, Craig K. (2003), 'Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data', *Psychological Methods* **8**(3), 322–337.

Faceli, Katti, Ana Carolina Lorena, João Gama & André C. P. L. F Carvalho (2011), *Inteligência Artificial. Uma abordagem de aprendizado de máquina*, 1th[a] edição, GEN.

Filatro, A. (2019), *Design instrucional 4.0*, Saraiva Educação S.A.
URL: *https://books.google.com.br/books?id=X_K7DwAAQBAJ*

Fleming, Philip J. & John J. Wallace (1986), 'How not to lie with statistics: The correct way to summarize benchmark results', *Commun. ACM* **29**(3), 218–221.
URL: *https://doi.org/10.1145/5666.5673*

Ghanem, A.S., S. Venkatesh & G. West (2008), Learning in imbalanced relational data.
URL: *https://www.scopus.com/inward/record.uri?eid=2-s2.0-77957947784partnerID=40md5=2817c8667170c25778328f295fc294a5*

Google (2020), 'Classification: Roc curve and auc', https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#roc-curve. Accessed: 2020-08-23.

Gray, Cameron C. & Dave Perkins (2019), 'Utilizing early engagement and machine learning to predict student outcomes', *Computers Education* **131**, 22 – 32.
URL: *http://www.sciencedirect.com/science/article/pii/S0360131518303191*

Grósz, Tamás, Gábor Gosztolya & László Tóth (2017), Training context-dependent dnn acoustic models using probabilistic sampling, *em* 'Proc. Interspeech 2017', pp. 1621–1625.
URL: *http://dx.doi.org/10.21437/Interspeech.2017-338*

Guyon, Isabelle, Jason Weston, Stephen Barnhill & Vladimir Vapnik (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1), 389–422.
URL: *https://doi.org/10.1023/A:1012487302797*

Haibo He, Yang Bai, E. A. Garcia & Shutao Li (2008), Adasyn: Adaptive synthetic sampling approach for imbalanced learning, *em* '2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)', pp. 1322–1328.

HAIR, JOSEPH F., BILL BLACK, BARRY BABIN, ROLPH E. ANDERSON & RONALD L. TATHAM (2009), *Análise Multivariada de Dados*, 6thª edição, bookman.

HAYKIN, S.S. (2001), *Redes Neurais - 2ed.*, BOOKMAN COMPANHIA ED.

He, H. & E. A. Garcia (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

He, Haibo & Yunqian Ma (2013), *Imbalanced Learning: Foundations, Algorithms, and Applications*, Wiley-IEEE Press.

Hintze, Jerry L. & Ray D. Nelson (1998), 'Violin plots: A box plot-density trace synergis', *The American Statistician* **52**(2), 181–184.

Hlosta, M., Z. Zdrahal & J. Zendulka (2017), Ouroboros: Early identification of at-risk students without models based on legacy data, pp. 6–15.

Huang, Shaobo & Ning Fang (2013), 'Predicting student academic performance in an engineering dynamics course : A comparison of four types of predictive mathematical models', *Computers & Education* **61**, 133–145.
URL: *http://dx.doi.org/10.1016/j.compedu.2012.08.015*

Igual, Laura & Santi Seguí (2017), *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*, 1thª edição, Springer.

Ippolito, P. P. (2019), 'Feature selection techniques', https://towardsdatascience.com/feature-selection-techniques-1bfab5fe0784. Accessed: 2020-01-20.

Izenman, A. J (2008), *Modern Multivariate Statistical Techniques*, Springer.

James, Gareth, Daniela Witten, Trevor Hastie & Robert Tibshirani (2014), *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated.

Jayaprakash, Sandeep M., Erik W. Moody, Eitel J.M. Lauría, James R. Regan & Joshua D. Baron (2014), 'Early alert of academically at-risk students: An open source analytics initiative', *Journal of Learning Analytics* pp. 6–47.

Johansson, Ulf, Cecilia Sönströd, Ulf Norinder & Henrik Boström (2011), 'Trade-off between accuracy and interpretability for predictive in silico modeling', *Future Medicinal Chemistry* **3**(6), 647–663. PMID: 21554073.
URL: *https://doi.org/10.4155/fmc.11.23*

Kahneman, D. & C. de Arantes Leite (2012), *Rápido e devagar: Duas formas de pensar*, Objetiva.
URL: *https://books.google.com.br/books?id=d3FloqhQHgQC*

Kanter, James Max & Kalyan Veeramachaneni (2015), Deep feature synthesis: Towards automating data science endeavors, *em* '2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Paris, France, October 19-21, 2015', IEEE, pp. 1–10.

Katrutsa, Alexandr & Vadim Strijov (2017), 'Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria', *Expert Systems with Applications* **76**, 1 – 11.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417417300635*

Kaya, Heysem & Alexey A. Karpov (2018), 'Efficient and effective strategies for cross-corpus acoustic emotion recognition', *Neurocomputing* **275**, 1028 – 1034.
URL: *http://www.sciencedirect.com/science/article/pii/S0925231217315680*

Kitchenham, Barbara & Stuart Charters (2007), Guidelines for performing systematic literature reviews in software engineering, Relatório Técnico EBSE 2007-001, Keele University and Durham University Joint Report.
URL: *http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8.pdf*

Kohavi, Ron (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *em* 'Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2', IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 1137–1143.

Kubat, Miroslav & Stan Matwin (1997), Addressing the curse of imbalanced training sets: One-sided selection, *em* 'In Proceedings of the Fourteenth International Conference on Machine Learning', Morgan Kaufmann, pp. 179–186.

L. G. F. Silva, M. E. P. S. Rocha & R. A. A. Fagundes (2017), 'Enade: Math and science students' performance analysis', *IEEE LATIN AMERICA TRANSACTIONS* **15**(09).

LAK (2011), 'Lak 2011 : 1st international conference learning analytics and knowledge', http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=11606. Accessed: 2020-01-12.

Lawrence, Steve, Ian Burns, Andrew Back, Ah Chung Tsoi & C. Lee Giles (2012), *Neural Network Classification and Prior Class Probabilities*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 295–309.
URL: *https://doi.org/10.1007/978-3-642-35289-8*$_1$9

Lemaître, Guillaume, Fernando Nogueira & Christos K. Aridas (2017), 'Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning', *Journal of Machine Learning Research* **18**(17), 1–5.
URL: *http://jmlr.org/papers/v18/16-365.html*

Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang & Huan Liu (2017), 'Feature selection: A data perspective', *ACM Comput. Surv.* **50**(6).
URL: *https://doi.org/10.1145/3136625*

Li, K. F., D. Rusk & F. Song (2013), Predicting student academic performance, *em* '2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems', pp. 27–33.

Little, Roderick J. A. & Nathaniel Schenker (1995), *Missing Data*, Springer US, Boston, MA, pp. 39–75.
URL: *https://doi.org/10.1007/978-1-4899-1292-3*$_2$

Liu, X. Y., J. Wu & Z. H. Zhou (2006), Exploratory under-sampling for class-imbalance learning, *em* 'ICDM '06: Proceedings of the Sixth International Conference on Data Mining,', pp. 965–969.

Lykourentzou, Ioanna, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis & Vassili Loumos (2009), 'Dropout prediction in e-learning courses through the combination of machine learning techniques', *Computers & Education* **53**(3), 950 – 965.
URL: *http://www.sciencedirect.com/science/article/pii/S0360131509001249*

Madley-Dowd, Paul, Rachael Hughes, Kate Tilling & Jon Heron (2019), 'The proportion of missing data should not be used to guide decisions on multiple imputation', *Journal of Clinical Epidemiology* **110**, 63 – 73.
URL: *http://www.sciencedirect.com/science/article/pii/S0895435618308710*

McCandless, David (2014), *Knowledge Is Beautiful*, 1th^a edição, Harper Design.

McNemar, Quinn (1947), 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika* **12**(2), 153–157.
URL: *http://dx.doi.org/10.1007/BF02295996*

Meier, Y., J. Xu, O. Atan & M. v. d. Schaar (2015), Personalized grade prediction: A data mining approach, *em* '2015 IEEE International Conference on Data Mining', pp. 907–912.

Miller, Tim (2017), 'Explanation in artificial intelligence: Insights from the social sciences', *CoRR* **abs/1706.07269**.
URL: *http://arxiv.org/abs/1706.07269*

Molnar, Christoph (2019), *Interpretable machine learning: A guide for making black box models explainable*.
URL: *https://christophm.github.io/interpretable-ml-book/*

Monllaó Olivé, D., D. Q. Huynh, M. Reynolds, M. Dougiamas & D. Wiese (2019), 'A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses', *IEEE Transactions on Learning Technologies* **12**(2), 171–183.

Mori, Toshiki & Naoshi Uchihira (2018), 'Balancing the trade-off between accuracy and interpretability in software defect prediction', *Empirical Software Engineering* **24**, 779–825.

Morocho-Cayamcela, Manuel Eugenio, Haeyoung Lee & Wansu Lim (2019), 'Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions', *IEEE Access* **7**, 137184–137206.

Márquez-Vera, C., A. Cano, C. Romero, A.Y.M. Noaman, H. Mousa Fardoun & S. Ventura (2016), 'Early dropout prediction using data mining: A case study with high school students', *Expert Systems* **33**(1), 107–124.

Márquez-Vera, C., A. Cano, C. Romero & S. Ventura (2013), 'Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data', *Applied Intelligence* **38**(3), 315–330.

Nadar, N. & R. Kamatchi (2018), 'A novel student risk identification model using machine learning approach', *International Journal of Advanced Computer Science and Applications* **9**(11), 305–309.
URL: *https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059023784partnerID=40md5=ffa11b4c3ec7a94cf7e29c9fe90b63d5*

Nelson, Karen J., Carole Quinn, Andrew Marrington & John A. Clarke (2012), 'Good practice for enhancing the engagement and success of commencing students', *Higher Education* pp. 83–96.

NIST/U.S.A (2013), 'Nist/sematech e-handbook of statistical methods', https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm. Accessed: 2020-01-17.

Paul, Ranjit Kumar (2008), Multicollinearity : Causes , effects and remedies.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Periwal, N. & K. Rana (2017), An empirical comparison of models for dropout prophecy in moocs, Vol. 2017-January, pp. 906–911.

Pinker, S. & L.T. Motta (2018), *O Novo Iluminismo: EM DEFESA DA RAZÃO, DA CIÊNCIA E DO HUMANISMO*, COMPANHIA DAS LETRAS.
URL: *https://books.google.com.br/books?id=qSw2uQEACAAJ*

Popper, K.R. (2004), *A lógica da pesquisa científica*, Cultrix.
URL: *https://books.google.com.br/books?id=MbGLmeMU3pMC*

Punlumjeak, W., S. Rugtanom, S. Jantarat & N. Rachburee (2017), 'Improving classification of imbalanced student dataset using ensemble method of voting, bagging, and adaboost with under-sampling technique', *Lecture Notes in Electrical Engineering* **449**, 27–34.

Ram, S., Y. Wang, F. Currim & S. Currim (2015), Using big data for predicting freshmen retention.
URL: *https://www.scopus.com/inward/record.uri?eid=2-s2.0-84964663145partnerID=40md5=3ec1d5ec8eb43f3c2328885d6c767e2a*

Ramentol, E., J. Madera & A. Rodríguez (2019), 'Early detection of possible undergraduate drop out using a new method based on probabilistic rough set theory', *Studies in Fuzziness and Soft Computing* **377**, 211–232.

Raschka, Sebastian (2018), 'Model evaluation, model selection, and algorithm selection in machine learning', *CoRR* **abs/1811.12808**.
URL: *http://arxiv.org/abs/1811.12808*

Ribeiro, Marco Tulio, Sameer Singh & Carlos Guestrin (2016), "why should i trust you?": Explaining the predictions of any classifier, *em* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 1135–1144.
URL: *https://doi.org/10.1145/2939672.2939778*

Rieger, S. A., R. Muraleedharan & R. P. Ramachandran (2014), Speech based emotion recognition using spectral feature extraction and an ensemble of knn classifiers, *em* 'The 9th International Symposium on Chinese Spoken Language Processing', pp. 589–593.

Romero, C. & S. Ventura (2010), 'Educational data mining: A review of the state of the art', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **40**(6), 601–618.

Romero, C. & S. Ventura (2019*a*), 'Guest editorial: Special issue on early prediction and supporting of learning performance', *IEEE Transactions on Learning Technologies* **12**(2), 145–147.

Romero, Cristóbal & Sebastian Ventura (2019*b*), 'Guest editorial: Special issue on early prediction and supporting of learning performance', *IEEE Transactions on Learning Technologies* **12**(2), 145 – 147.
URL: *https://ieeexplore.ieee.org/document/8735954*

Rovira, Sergi, Eloi Puertas & Laura Igual (2017), 'Data-driven system to predict academic grades and dropout', *PLoS ONE* pp. 1–21.

Salvato, Marcio Antonio, Pedro Cavalcanti Gomes Ferreira & Angelo JosÃ© Mont'Alverne Duarte (2010), 'O impacto da escolaridade sobre a distribuiÃ§Ãde renda', *Estudos EconÃ´micos (SÃPaulo)* **40**, 753 – 791.
URL: *http://www.scielo.br/scielo.php?script=sci*$_a$*rttextpid* = *S*0101 − 41612010000400001*nrm* = *iso*

Samuel, A. L. (1959), 'Some studies in machine learning using the game of checkers', *IBM J. Res. Dev.* **3**(3), 210–229.
URL: *https://doi.org/10.1147/rd.33.0210*

Schuller, Björn W., Stefan Steidl & Anton Batliner (2009), The interspeech 2009 emotion challenge, *em* 'INTERSPEECH'.

Shahiri, Amirah Mohamed, Wahidah Husain & Nur'aini Abdul Rashid (2015), 'A review on predicting student's performance using data mining techniques', *Procedia Computer Science* **72**, 414 – 422. The Third Information Systems International Conference 2015.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050915036182*

Shavlik, J.W., T. Dietterich & T.G. Dietterich (1990), *Readings in Machine Learning*, Machine Learning Series, Morgan Kaufmann Publishers.
URL: *https://books.google.com.br/books?id=UgC33U2KMCsC*

Swalin, A. (2018), 'How to handle missing data', https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4. Accessed: 2020-01-20.

Swets, J A (1988), 'Measuring the accuracy of diagnostic systems', *Science* pp. 1285–93.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/3287615*

Thammasiri, Dech, Dursun Delen, Phayung Meesad & Nihat Kasap (2014), 'A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition', *Expert Systems with Applications* **41**(2), 321 – 330.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417413005332*

Ting, Kai Ming (2017), *Confusion Matrix*, Springer US, Boston, MA, pp. 260–260.
URL: *https://doi.org/10.1007/978-1-4899-7687-1_50*

Tinto, Vincent (1975), 'Dropout from higher education: A theoretical synthesis of recent research', *Review of Educational Research* pp. 89 – 125.

Tinto, Vincent (1982), 'Defining dropout: A matter of perspective', *New Directions for Institutional Research* .

Tinto, Vincent (2017), 'Through the eyes of students', *Journal of College Student Retention: Research, Theory & Practice* **19**(3), 254–269.
URL: *https://doi.org/10.1177/1521025115621917*

van der Maaten, Laurens & Geoffrey Hinton (2008), 'Visualizing data using t-SNE', *Journal of Machine Learning Research* **9**, 2579–2605.
URL: *http://www.jmlr.org/papers/v9/vandermaaten08a.html*

Verleysen, Michel & Damien François (2005), The curse of dimensionality in data mining and time series prediction, *em* J.Cabestany, A.Prieto & F.Sandoval, eds., 'Computational Intelligence and Bioinspired Systems', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 758–770.

Wang, Rui, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton & Andrew T. Campbell (2018), 'Tracking depression dynamics in college students using mobile phone and wearable sensing', *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**(1).
URL: *https://doi.org/10.1145/3191775*

Wattenberg, Martin, Fernanda Viégas & Ian Johnson (2016), 'How to use t-sne effectively', *Distill* .
URL: *http://distill.pub/2016/misread-tsne*

WILK, M. B. & R. GNANADESIKAN (1968), 'Probability plotting methods for the analysis of data', *Biometrika* **55**(1), 1.

Witten, Ian H., Eibe Frank, Mark A. Hall & Christopher J. Pal (2016), *Data Mining Practical Machine Learning Tools and Techniques*, Elsevier.

Wolpert, David H. (2002), *The Supervised Learning No-Free-Lunch Theorems*, Springer London, London, pp. 25–42.
URL: *https://doi.org/10.1007/978-1-4471-0123-9*$_3$

Xu, J., K. H. Moon & M. van der Schaar (2017), 'A machine learning approach for tracking and predicting student performance in degree programs', *IEEE Journal of Selected Topics in Signal Processing* **PP**(99), 1–1.

Y. Amaya, E. Barrientos & D. Heredia (2015), 'Student dropout predictive model using data mining techniques', *IEEE LATIN AMERICA TRANSACTIONS* **13**(09).

Zhu, Xingquan & Ian Davidson (2007), *Knowledge Discovery and Data Mining: Challenges and Realities*.