Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Perfil de Evasão no Ensino Superior Brasileiro: uma Abordagem de Mineração de Dados

Lucas Rocha Soares de Assis

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Guilherme Novaes Ramos

Brasília
2017

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Rp

Rocha Soares de Assis, Lucas
Perfil de Evasão no Ensino Superior Brasileiro: uma Abordagem de Mineração de Dados / Lucas Rocha Soares de Assis; orientador Guilherme Novaes Ramos. -- Brasília, 2017.
153 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2017.

1. Evasão no ensino superior. 2. Mineração de dados. 3. Aprendizagem de máquina. 4. Classificação. I. Novaes Ramos, Guilherme, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Perfil de Evasão no Ensino Superior Brasileiro: uma Abordagem de Mineração de Dados

Lucas Rocha Soares de Assis

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Guilherme Novaes Ramos (Orientador)
CIC/UnB

Prof. Dr. Donald Matthew Pianto
Universidade de Brasília

Prof. Dr. Remis Balaniuk
Universidade Católica de Brasília

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 12 de dezembro de 2017

Dedicatória

Aos meus pais, irmão e esposa.

Agradecimentos

Agradeço a todos que contribuíram direta ou indiretamente para a realização deste trabalho. Em especial, ao Professor Guilherme, pela orientação, conselhos e ensinamentos. À equipe de docentes do Programa de Mestrado em Computação Aplicada, ao Coordenador Marcelo Ladeira, pela oportunidade de aprendizado e crescimento profissional. Aos professores Donald e Remis, que aceitaram compor a banca de defesa, pelas sugestões e análises significativas.

Agradeço também a meus pais e irmão, que me ensinaram valores e amor. À minha esposa, Pâmella, pela paciência e companheirismo. Aos colegas de mestrado, em especial ao Thiago, companheiro de mestrado, de trabalho e amigo. Aos meus colegas de trabalho, em especial, Laura, Katia, Nara, Douglas, Viviane e Janaina, pelo convívio diário, troca de experiências e por todo apoio necessário. Aos meus amigos, pela compreensão e incentivo.

Resumo

A evasão no ensino superior é um problema que atinge diversas instituições no mundo. No Brasil, não há divulgação regular de dados sobre o assunto. Neste trabalho foram aplicadas técnicas de mineração de dados para criar um perfil de estudantes que evadem do ensino superior brasileiro. Utilizando dados do Censo da Educação Superior (CES) e Exame Nacional do Ensino Médio (ENEM), foram criados modelos de cinco algoritmos de classificação para analisar a evasão de alunos ingressantes em três diferentes níveis de evasão. Os experimentos foram conduzidos nos dados de estudantes da UnB. Entre os algoritmos testados, o CART obteve um desempenho marginalmente superior, na métrica de sensibilidade. Ele obteve desempenho de cerca de 84% para evasão a nível de curso. Nos demais testes, não houve diferença estatisticamente significativa entre os algoritmos. As principais características identificadas nos alunos que possuem propensão a evadir são: ingressar no primeiro semestre, possuir vínculos com mais de uma IES, obter notas acima da média nos exames do ENEM e já ter concluído o ensino médio no momento que realiza as provas do ENEM. Também foi desenvolvido um pacote para o R em que é possível treinar novos classificadores de evasão, que podem ser utilizados para determinar, em qualquer IES ou grupo de IES, quais alunos possuem maior tendência de evadir.

Palavras-chave: evasão no ensino superior, mineração de dados, aprendizagem de máquina, classificação

Abstract

Student attrition and eventual dropout are problems that affect many universities around the world. In Brazil, there are no official statistics to monitor them. In this work, data mining techniques were used in order to unveil a profile of dropout students from the Brazilian higher education. The data from Brazil's higher education census, CES, and its nation-wide high school exam, ENEM, were used to create multiple classification models, ranging from five different classification methods and three separate dropout definitions. The experiments were conducted on UnB's students' data. Among the classification methods, CART showed a subtle lead, performance wise. It obtained a sensibility score of around 84% when the dropout definition was focused on the student's major. On the other two dropout definitions, there wasn't a statistically significant difference between the tested methods. The main characteristics for the dropout students unveiled by the generated models were: to enter the university in the first semester, attend to more than one institution, obtain higher than average grades on the high school examinations and finally, having graduated from high school when taking the ENEM exam. Furthermore, a R package was developed in order to train new classifiers for dropout. It can be used to determine, in a given database, which students are more likely to dropout.

Keywords: higher education attrition, dropout, data mining, machine learning, classification

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABI Área Básica de Ingresso.

CES Censo da Educação Superior.

CP *Complexity Parameter*.

CPF Cadastro de Pessoa Física.

CRISP-DM *CRoss-Industry Standard Process for Data Mining*.

ENADE Exame Nacional de Desempenho de Estudantes.

ENEM Exame Nacional do Ensino Médio.

FIES Fundo de Financiamento ao Estudante do Ensino Superior.

IBGE Instituto Brasileiro de Geografia e Estatística.

IES Instituições de Ensino Superior.

Inep Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

ISCED *International Standard Classification of Education*.

KDD *Knowledge Discovery in Databases*.

MEC Ministério da Educação.

ProUni Programa Universidade para Todos.

SEEC Serviço de Estatística da Educação e Cultura.

SISU Sistema de Seleção Unificada.

SVM *Support Vector Machines*.

UnB Universidade de Brasília.

Capítulo 1

Introdução

A evasão, entendida como uma interrupção no ciclo de estudos, é um problema que atinge instituições de ensino em geral [1]. Este é um tópico de crescente interesse ao redor do mundo [2] e está entre as principais agendas políticas de alguns países, como Espanha e Itália [3]. No Brasil, o debate sobre a evasão na educação básica está melhor estabelecido que o debate sobre a evasão no ensino superior [4]. Os estudos brasileiros sobre evasão no ensino superior ainda são incipientes e apenas começaram a crescer em quantidade na década de 2000 [5], enquanto nos Estados Unidos, por exemplo, acontecem desde a década de 1960 [6].

1.1 Evasão no ensino superior

No Brasil, os levantamentos de dados sobre a educação superior são realizados desde meados do século passado, mas, inicialmente, sem planejamento ou periodicidade. Em 1956, com a criação do Serviço de Estatística da Educação e Cultura (SEEC) e cooperação do Instituto Brasileiro de Geografia e Estatística (IBGE), a coleta é sistematizada e passa a ser feita anualmente [7]. No início da década de 1980, o SEEC é realocado do Rio de Janeiro para Brasília e passa a integrar a secretaria de informática do Ministério da Educação (MEC) [7]. Em 1997, a lei nº 9.448 [8], que transforma o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) em Autarquia Federal, estabelece que o Inep é responsável por coletar, analisar e disseminar informações sobre a educação. É criado então o Censo da Educação Superior (CES), pesquisa realizada pelo Inep com o objetivo de coletar informações sobre todas Instituições de Ensino Superior (IES) do Brasil [9]. Todas IES são obrigadas, por vários dispositivos legais [8, 10, 11], a informar ao CES a relação dos dados sobre seus corpos docente e discente. A partir de 2010, o CES passou a coletar informações individualizadas dos estudantes como Cadastro de Pessoa Física (CPF), situação do aluno no curso, data de ingresso [7], entre outras.

Apesar da divulgação sistemática de dados sobre o ensino superior brasileiro ao nível de alunos pelo Inep[1], não há divulgação oficial consolidada de informações ou de indicadores sobre a evasão. São duas as características da evasão que dificultam a formulação de estatísticas e indicadores. Primeiramente, o contexto do aluno no ensino superior é complexo, pois sua trajetória pode tomar diversos caminhos [12]. Um estudante pode estar vinculado a diversas instituições e diversos cursos ao mesmo tempo, podendo desistir de apenas um dos cursos, transferir-se entre cursos ou instituições e formar-se em um dos cursos. Ademais, não há definição generalizada de evasão, pois esse é um problema que depende do contexto estudado [6, 12]. Por exemplo, em uma IES a evasão pode ser definida de forma que apenas os alunos jubilados são considerados. Nesse caso, alunos desistentes não seriam considerados alunos evadidos.

A definição de evasão no ensino superior também pode ser feita em níveis. Por exemplo, um estudante pode deixar de cursar uma determinada disciplina, mas não abandonar o curso. Neste caso, para o coordenador do curso, o estudante pode ser considerado uma evasão na disciplina, mas não no curso. De forma similar, um estudante pode também abandonar totalmente um determinado curso, mas sem deixar a instituição, sendo considerado uma evasão apenas no curso. Um estudante pode também ser rotulado como evasão de uma instituição, ou do conjunto de todas instituições do país, ou subconjuntos do país como, categorias administrativas (pública ou privada), formas de organização acadêmica (universidades, faculdades, centros universitários, entre outros) ou até áreas de conhecimento [13].

No Brasil, grande parte dos estudos a respeito da evasão no ensino superior tem origem nas IES [14]. Esses estudos concentram-se nos problemas de evasão das instituições, geralmente de algum curso específico que possui uma elevada taxa de evasão. Embora existam estudos sobre a evasão no país como um todo [4, 13], não são realizadas análises ao nível do aluno. Desde 2010 com os dados coletados pelo Inep, é possível realizar análises a nível do aluno para qualquer IES do Brasil.

A evasão gera alguns problemas, como a perda de recursos pelo Estado quando um aluno ingressa em uma de suas instituições públicas de ensino superior e falha em adquirir um diploma. Não só há um desperdício de recursos financeiros por parte da instituição, como também prejuízo social para todo o país ao não qualificar sua mão de obra [13]. O mesmo prejuízo social ocorre nas instituições privadas, que são maioria no Brasil [4]. Ocorre também um prejuízo individual, sofrido pelos estudantes que evadem, eles despendem tempo e dinheiro e não obtêm a qualificação que almejavam [6, 12]. No Brasil, informações a respeito do fenômeno da evasão são escassas o que torna difícil a identificação de pontos críticos do problema. Em outros países, como Espanha, Itália e Estados

[1] `http://www.inep.gov.br/`

Unidos, a evasão é estudada a ponto de se saber quais são as principais características que levam os alunos a deixar suas IES, possibilitando a criação de políticas públicas específicas ou ações das instituições para limitar o problema [2, 3, 15]. No Brasil, nem mesmo indicadores de retenção dos alunos em seus primeiros anos de estudo são divulgados regularmente.

1.2 Questões de pesquisa

Um dos problemas enfrentados no Brasil em relação à evasão é a pouca quantidade de informações sobre o assunto, como a falta de indicadores de evasão. Para começar a resolver esse problema se faz necessário definir o que é evasão nas bases de dados do Inep. Portanto, a primeira questão que direciona este trabalho é: como reconhecer quais alunos são evadidos e quais são retenção nas bases de dados do Inep, mais precisamente da base do CES?

Uma vez identificados os alunos evadidos e retidos (não evasão), duas questões são levantadas: é possível identificar os fatores envolvidos na evasão desses alunos? Se sim, quais seriam esses fatores? Respondendo positivamente às questões levantadas, uma última questão é colocada: como auxiliar às IES a entender o problema e por consequência diminuir suas taxas de evasão?

A complexidade do problema e a grande quantidade de dados disponível nas bases do Inep, demandam o uso de técnicas capazes de extrair informações potencialmente úteis da massa de dados. Por essa razão, mineração de dados será fundamental no desenvolvimento deste trabalho.

1.3 Justificativa

O nível educacional de uma nação influencia sua economia de diversas formas [16]. Em geral, existe uma correlação positiva entre o nível de escolarização de um país, sua produtividade e os ganhos salariais de sua população [16, 17]. Portanto, a evasão no ensino superior pode influenciar negativamente alguns aspectos da economia de um país.

Ademais, uma característica da população brasileira pode agravar o potencial impacto econômico. O gradual envelhecimento dessa população [18] e a diminuição no crescimento dos ingressos do ensino superior nos anos de 2012, 2013 e 2014 [19] indicam que o ensino superior deverá contar com menos ingressantes com o passar dos anos. Portanto, mantendo-se a atual proporção de formados por ingressantes, a diminuição de ingressantes nas IES acarretará em um menor número de formados no futuro. Uma das formas de limitar a queda no número de formados é através da diminuição da evasão.

A fim de diminuir a evasão no Brasil é necessário primeiro entender seu comportamento. A falta de informação acerca do fenômeno da evasão no ensino superior no país é um problema que precisa ser remediado. O Inep é uma autarquia que possui dentro de suas finalidades a elaboração de diagnósticos a respeito da educação [10]. No entanto, não há dentro do Inep estudo desenvolvido em busca de descrever as características dos alunos evadidos do ensino superior brasileiro.

1.4 Contribuições esperadas

No presente trabalho, a partir dos resultados de diferentes algoritmos de classificação, será conduzida uma análise buscando discernir quais características dos alunos impactam no fenômeno da evasão em cada um dos cenários estudados. Pretende-se, ainda, criar um *framework* para gerar classificadores de alunos de graduação presencial em diferentes cenários. Esse *framework* será disponibilizado para que, com as coletas de dados subsequentes feitas pelo Inep, eles possam servir para indicar quais alunos estão em maior risco de evadir e consequentemente apoiar medidas que possam diminuir a taxa de evasão. Espera-se que o trabalho contribua para um melhor entendimento do problema da evasão no país com essas contribuições acadêmicas e tecnológicas.

1.5 Objetivos

O objetivo geral deste trabalho é identificar as principais características de alunos que evadem do ensino superior brasileiro através do uso de técnicas de mineração de dados, com cinco algoritmos de classificação.

Os objetivos específicos são:

- verificar se é possível identificar alunos evadidos utilizando os dados do Inep em uma IES;
- verificar, dentro do grupo de algoritmos de classificação selecionados, qual produz resultados mais promissores em termos das métricas utilizadas;
- analisar casos específicos de grupos de alunos classificados como evasão; e
- disponibilizar um *framework* para treinar diferentes classificadores em grupos de dados específicos.

Capítulo 2

Revisão de Literatura

Este capítulo aborda uma revisão de literatura a respeito da evasão no ensino superior, trabalhos relacionados, que classificam estudantes como em risco de evasão em universidades, apresenta também uma breve introdução à mineração de dados e aprendizado de máquina e faz uma revisão de técnicas de classificação bem sucedidas ao classificar estudantes em risco de evasão.

2.1 Evasão no ensino superior

Muitos modelos teóricos foram desenvolvidos com o propósito de conhecer as características dos estudantes que evadem do ensino superior [20, 21, 22]. O modelo teórico mais comumente citado na literatura [12, 23, 24] é o desenvolvido por Tinto [6]. Seu modelo descreve o processo de evasão de um aluno como sendo determinado por atributos individuais, atributos familiares, habilidades anteriores à entrada no ensino superior, integração social e acadêmica dentro da universidade, comprometimento individual, comprometimento da instituição e fatores sociais e familiares externos ao aprendizado acadêmico. Tradicionalmente, estudos de evasão de estudantes do ensino superior têm sido realizados através de pesquisas longitudinais e qualitativas que acompanham uma coorte[1] de estudantes por um período predeterminado, as chamadas *surveys* [23]. Em geral, as informações para o estudo são adquiridas através do preenchimento de questionários que são apresentados periodicamente aos estudantes da coorte [20, 23]. Apesar desses estudos terem formado a base teórica para os estudos de evasão no ensino superior [3, 20, 23], algumas críticas são pertinentes: a falta de generalização da pesquisa limita a aplicabilidade de seus resultados em diferentes instituições; a dificuldade e alto custo da implementação de questionários em larga escala em longos períodos de tempo; amostras de tamanhos reduzidos são realizadas

[1]Conjunto de pessoas que têm um evento em comum. Exemplo: estudantes que ingressaram na Universidade de Brasília (UnB) em 2016.

devido ao alto custo de coleta de dados, o que pode comprometer a representatividade dos estudos. No caso brasileiro, Gaioso [1] alega que a teoria de Tinto pode não ser apropriada, em função da deficiência da educação básica e da situação sócio-econômica da população brasileira, que é distinta da norte americana. Uma alternativa a esses estudos *survey* é uma abordagem analítica e quantitativa, em que as informações comumente encontradas nas bases de dados das instituições são utilizadas [23].

2.2 Trabalhos relacionados

O trabalho de Sarker *et al.* [20] realiza um estudo de classificação de alunos em risco de evasão na Southampton University na Inglaterra. Os autores utilizam dados do registro interno da universidade, de questionários e de dados abertos e ligados disponíveis no país. Busca-se mostrar que os dados abertos ligados, que possuem questionários feitos a alunos de toda Inglaterra podem substituir os modelos tradicionais de questionários aplicados aos alunos investigados. São utilizados redes neurais com *multilayer perceptron* e regressão logística para classificar os alunos em risco de evadir no primeiro ano de estudo. Ambos os classificadores obtém resultados muito próximos em acurácia, sensibilidade e sensitividade. Sendo a regressão logística eleita como o melhor classificador, pois sua sensibilidade é maior que no modelo de redes neurais. A sensibilidade é importante nesse tipo de estudo, pois falsos-negativos, ou alunos em risco de evasão sendo classificados erroneamente, devem ser evitados. Isto é, um resultado falso-positivo é menos prejudicial nesses casos, pois dar uma atenção especial para um aluno que provavelmente não iria evadir é menos prejudicial que não dar a atenção devida a um aluno que necessita de assistência para não deixar os estudos.

Em [23], Delen busca classificar se os estudantes estão em risco de evasão ou não dentro do primeiro ano de universidade. O autor utilizou uma base de dados de uma universidade dos Estados Unidos que continha 8 anos de informação de primeiro anistas. Entre os atributos utilizados estavam as notas dos estudantes no primeiro semestre dentro da universidade, assim como, notas do exame de ensino médio americano, o SAT[2]. O autor utilizou validação cruzada de tamanho 10 e comparou três algoritmos para classificar os alunos em risco. Redes neurais com *multilayer perceptron* e *back propagation*, C5.0 e regressão logística. Dentre os três, a rede neural obteve as melhores médias gerais em acurácia, sensibilidade e sensitividade, com valores muito próximos ao algoritmo de regras de decisão (C5.0).

Dekker *et al.* [25], buscam classificar alunos em risco de evadirem no primeiro ano de estudo de um curso de engenharia na Holanda. Primeiramente, classificam-se os estudan-

[2]É o exame que avalia o desempenho de alunos do ensino médio, similar ao ENEM.

tes utilizando apenas uma base de dados que contém informações antes do ingresso na instituição. Posteriormente, os autores incorporam às informações iniciais, informações do rendimento dos alunos dentro da instituição. São utilizados vários classificadores que segundo o autor, são populares na ferramenta Weka: *OneR* (algoritmo que utiliza apenas uma regra de decisão), CART, C4.5, *BayesNet*, regressão logística, *JRip* (algoritmo de aprendizado de regras) e *RandomForest*. Todos os modelos obtêm desempenho superior quando classificam os alunos utilizando as informações completas, ou seja, informações antes do ingresso e informações durante o curso. Além disso, os algoritmos que obtêm os melhores desempenhos são o CART e o C4.5, embora a diferença para regressão logística não tenha sido estatisticamente significativa.

Em [26], Zhang *et al.* conduzem um estudo semelhante a [20]. A proposta é classificar os estudantes que estão em risco de evadir da Thames Valley University na Inglaterra. Os autores utilizam três algoritmos para classificar os estudantes. *Naive Bayes*, *Support Vector Machines* (SVM) e ID3, um algoritmo de árvore de decisão. O *Naive Bayes* é o que obtém a melhor acurácia e sensibilidade. Os autores verificam que os atributos mais importantes são as notas que os estudantes atingem dentro da universidade e o uso da biblioteca virtual. O atributo que menos informação agrega para explicar a evasão é a cor/raça do estudante. O estudo não utiliza informações prévias a entrada dos estudantes no ensino superior e apenas informações relativas a um ano da universidade são utilizadas.

Não há um consenso claro na literatura sobre qual o melhor algoritmo para se classificar alunos em risco de evasão [25, 27]. Na maior parte dos casos, algoritmos de regras de decisão, regressão logística e redes neurais obtiveram mais sucesso em classificar alunos em risco de evasão [20, 23, 25, 28, 29].

2.3 Aprendizagem de máquina

O nome "aprendizagem de máquina" foi cunhado em 1980 durante os primeiros encontros de pesquisadores interessados em abordagens computacionais de aprendizagem [30]. O campo é derivado da inteligência artificial (ciência da computação) e ciência cognitiva quando, à época, essas ciências mostravam pouco interesse em estudar problemas relacionados à aprendizagem, preferindo focar no papel que o conhecimento exerce na inteligência [30]. A definição do que é "aprender" é um conceito filosófico. Afinal, máquinas podem aprender? Uma definição mais assertiva e ligada a mineração de dados e aprendizagem de máquina seria a de que quando algo muda seu comportamento de uma forma a melhorar seu desempenho futuro, houve aprendizado [31].

Aprendizagem de máquina estendeu-se para diversas áreas do conhecimento e seu impacto é percebido no dia-a-dia da sociedade com aplicações como filtro de *spam* em

e-mails; ferramentas de busca na internet; sistemas de recomendação para compras e desenvolvimento de novos medicamentos [32]. Sistemas de aprendizagem de máquina podem aprender programas automaticamente dos dados. Essa é uma alternativa atrativa em relação a construir programas complexos manualmente, em que a falta de percepção do todo e o tempo demandados são pontos negativos [33].

Basicamente, quatro diferentes tipos de aprendizagem são utilizados em aprendizagem de máquina: aprendizagem por classificação, aprendizagem por associação, aprendizagem por agrupamento e predição numérica [34]. Classificação é a técnica em que é apresentada à máquina uma coleção de exemplos previamente rotulados e espera-se que uma maneira de classificar novos exemplos seja aprendida. Associação é feita quando deseja-se obter qualquer associação entre atributos e não apenas aqueles que predizem uma classe específica. Em agrupamento, busca-se grupos que contenham exemplos similares. Predição numérica funciona como classificação, no entanto as classes não são discretas, mas contínuas [31].

2.4 Mineração de dados

Em diversas áreas do conhecimento, dados têm sido coletados e armazenados em um ritmo impressionante [27]. O grande aumento na coleta e armazenamento de dados criou uma necessidade de técnicas capazes de traduzirem ou extraírem informações potencialmente úteis da massa de dados. Historicamente, a noção de encontrar novos padrões úteis em bases de dados recebeu diversos nomes, como mineração de dados, *Knowledge Discovery in Databases* (KDD), extração de conhecimento, colheita de dados entre outros [35]. Apesar dos diferentes nomes e, em alguns casos abordagens, os termos mencionados tem como objetivo principal a extração de conhecimento de bases de dados, sendo, portanto identificados neste trabalho como mineração de dados.

Mineração de Dados foi formada e evolui a partir da interseção de diferentes áreas do conhecimento, como estatística e ciência da computação [36]. Técnicas de mineração de dados têm sido amplamente utilizadas para resolver problemas de diversas áreas como educação [27], indústria, ciência, engenharia [37] e Estado [38]. Mineração de dados pode ser compreendida como o processo não trivial de se identificar padrões válidos, novos, potencialmente úteis e compreensíveis na massa de dados. Além disso, mineração de dados possui forte ênfase em se trabalhar com grandes quantidades de dados reais. Portanto, a escalabilidade dos algoritmos utilizados é de fundamental interesse [35].

A busca por procedimentos e técnicas que extraíssem conhecimento das massas de dados incentivou acadêmicos e a indústria a criarem padrões para eles [39]. Os esforços para estabelecer um padrão foram iniciados na academia, com o desenvolvimento de dois

modelos de processos [40]. O modelo de nove etapas de Fayyad *et al.* [35] e o modelo de oito etapas de Anand e Buchner [32]. A ênfase dos modelos era fornecer uma sequência de atividades que ajudasse a executar um projeto de mineração de dados em um domínio arbitrário [41]. Em seguida, modelos de processos direcionados para a indústria foram desenvolvidos. Das diferentes abordagens propostas, duas se destacaram, a proposta de cinco etapas de Cabena *et al.* [42] e o *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) [43], desenvolvido por um consórcio de empresas européias e que possui seis etapas.

2.5 Classificação

Classificação é a técnica mais madura em aprendizagem de máquina e também uma das mais utilizadas [33]. Em problemas de classificação, um algoritmo consome uma coleção de instâncias como entrada, cada uma pertencendo a uma classe e cada uma contendo um grupo de atributos. O sistema é então treinado em parte dos dados e posteriormente tenta identificar a qual classe uma nova instância, previamente não vista, pertence [34]. Classificação pode ser vista como uma técnica supervisionada de aprendizagem, pois geralmente as classes são rotuladas e fornecidas ao sistema para o seu treinamento [31].

2.5.1 Regressão logística

A regressão logística, assim como a regressão linear, é uma técnica estatística que estuda a relação entre uma variável resposta ou classe e uma ou mais variáveis independentes ou atributos com o objetivo de obter o melhor modelo possível que seja capaz de predizer a variável resposta. Por melhor possível, entende-se a maximização da função de verossimilhança [44].

A diferença entre regressão linear e regressão logística está no fato de que a variável resposta da regressão logística é discreta ou categórica, sendo na maioria das vezes dicotômica (sucesso/fracasso). Outra característica da regressão logística é a capacidade de estimar a probabilidade p_i de sucesso e, consequentemente, a probabilidade $1 - p_i$ de fracasso, dados determinados atributos do indivíduo avaliado [23]. Assim como algoritmos de árvores de decisão, regressão logística pode ser uma boa escolha quando uma interpretação do modelo se faz necessária, pois os coeficientes gerados para cada atributo indicam seu impacto na classificação da classe de estudo [44]. A regressão logística assume que existe uma relação linear entre os atributos preditores e a razão de chance da classe estudada o que limita sua capacidade de separar classes que se relacionam com os atributos de forma não linear. Ademais, regressão logística exige que o usuário determine a relação entre os atributos preditores e a classe estudada [31].

Um modelo de regressão logística pode ser descrito por:

$$E(Y|x) = G(X) = G(\beta_0 + \beta_i X_i) \tag{2.1}$$

Em que, $G(X)$ é uma função cujos valores variam estritamente entre 0 e 1, ($0 < G(X) < 1$), Y é a variável resposta ou a classe, β é o vetor dos coeficientes desconhecidos das variáveis independentes ou atributos utilizados no modelo e X é o vetor das instâncias. Apesar de existirem várias transformações para garantir que $G(X)$ seja uma probabilidade, as mais utilizadas são o *logit* e o *probit* [45].

A probabilidade de sucesso/insucesso é obtida a partir do modelo:

$$P(Y = 1|x) = G(\beta X) \tag{2.2}$$

$$P(Y = 0|x) = 1 - G(\beta X) \tag{2.3}$$

2.5.2 Árvores de decisão

Árvores de decisão são baseadas em uma abordagem de divisão e conquista aplicada ao problema de classificação [34]. O problema de construção de uma árvore de decisão pode ser expressa recursivamente. Um atributo é selecionado para servir de raiz da árvore e ramos são criados a partir de cada valor do atributo selecionado. Isso divide as instâncias em subgrupos, um para cada valor do atributo escolhido. Em seguida, o processo é repetido para cada um dos subgrupos, utilizando apenas as instâncias que alcançam aquele ramo. Quando todas as instâncias de um nó possuem a mesma classe, a árvore para de crescer para aquele nó [31].

Seja S um conjunto de instâncias a serem classificadas, cada uma pertencendo a uma classe C_i, $freq(C_i, S)$ a frequência em que cada item de S pertence a classe C_i e $\#S$ a quantidade de instâncias no conjunto S. Ao selecionar uma instância aleatoriamente e declarar que ela pertence a alguma classe C_i, tem-se que essa mensagem possui probabilidade de:

$$P(C) = \frac{freq(C_i, S)}{\#S} \tag{2.4}$$

A medida de informação trazida pela mensagem depende de sua probabilidade. Ela é medida em *bits* e é dada por:

$$I(S) = -\log_2 \left(\frac{freq(C_i, S)}{\#S} \right) bits \tag{2.5}$$

Para encontrar a informação esperada de tal mensagem, soma-se através de todas as classes, proporcionalmente às suas frequências em S:

$$H(S) = -\sum_{j=1}^{k} \frac{freq(C_i, S)}{\#S} \times I(S) \tag{2.6}$$

$H(S)$ também é conhecida como a entropia do conjunto S ou quando aplicado a um conjunto de treinamento T, $H(T)$ mede a informação média necessária para identificar a classe de uma dada instância em T. Após o conjunto T ser dividido de acordo com os n resultados da escolha de um atributo X, a informação esperada necessária pode ser descrita como:

$$H_X(T) = \sum_{i=1}^{n} \frac{\#T_i}{\#T} \times H(T_i) \tag{2.7}$$

O ganho de informação ao se dividir o conjunto de treinamento T de acordo com o atributo X é dado pela diferença entre $H(T)$ e $H_X(T)$. O critério de ganho é realizado então, quando seleciona-se o atributo que maximiza o ganho de informação [46].

Os C4.5 e C5.0, desenvolvidos por Ross Quinlan, são implementações de árvores de decisão. Para a seleção do atributo que servirá de raiz e eventualmente as escolhas dos subgrupos, ambos utilizam o critério de ganho de informação [46]. Quinlan afirma que o algoritmo C5.0 é superior ao C4.5 [47], no entanto, o C5.0 não possui documentação própria [48]. O que existe é sua implementação na linguagem C[3] e um documento feito por Quinlan que descreve as mudanças em relação ao C4.5 [49], que possui documentação própria [46]. A partir desses insumos, foi implementado o C50, pacote para a linguagem R [50] que treina árvores e regras de decisão com o algoritmo C5.0 no R.

Algumas diferenças do C5.0 em relação ao C4.5 é a implementação de *boosting*, aumento da velocidade de processamento com a criação de árvores menores que as criadas pelo C4.5 e *winnowing*, a capacidade do algoritmo de ignorar atributos que são apenas marginalmente relevantes. *Boosting* é uma técnica que combina vários modelos para criar um modelo melhor. São atribuídos pesos às observações de treinamento, e esses são variados para que cada novo modelo tenha foco maior nos erros de classificação dos modelos anteriores [31].

Nas implementações de C4.5 e C5.0 por Ross Quinlan, um critério de ganho modificado é utilizado. O critério de ganho descrito anteriormente possui uma deficiência, ele favorece atributos que possuem muitos valores distintos, por exemplo, um código identificador das instâncias [46]. Para contornar essa deficiência, uma normalização é realizada:

[3] `http://rulequest.com/GPL/C50.tgz`

$$SplitH_X(T) = \sum_{i=1}^{n} \frac{\#T_i}{\#T} \times \log_2 \left(\frac{\#T_i}{\#T} \right) \tag{2.8}$$

Ajustando o ganho de informação pelos valores distintos que o atributo possui, tem-se:

$$Ganho_{adj}(T_X) = (H(T) - H_X(T))/SplitH_X(T) \tag{2.9}$$

Outro algoritmo que cria árvores de decisão para classificação é o CART [31]. CART foi desenvolvido por um grupo de estatísticos em 1984 [51]. Diferentemente do C4.5 e C5.0, o CART utiliza a impureza de Gini para escolher os atributos em que a árvore se dividirá [51]. A impureza de Gini é uma medida de quão frequente uma instância, pertencente a um conjunto X, escolhida ao acaso seria incorretamente classificada se a distribuição dos rótulos fosse usada para classificá-la. Ela alcança zero quando todas as instâncias em um nó possuem a mesma classe. A medida pode ser computada através da Equação 2.10, onde j é o número total de classes e f_i a frequência de itens rotulados com a classe i.

$$I_G(f) = \sum_{i=1}^{j} f_i(1 - f_i) \tag{2.10}$$

Algoritmos de árvores de decisão podem criar árvores complexas que acabam se ajustando muito bem ao conjunto de treinamento, podendo perder a generalização que é fundamental para bons classificadores [52]. Por essa razão, ambos algoritmos, C5.0 e CART, possuem uma abordagem em que as árvores são podadas após a sua construção [46, 51]. Uma alternativa seria podar a árvore durante seu crescimento, o que poderia economizar tempo, já que partes das árvores não seriam construídas. No entanto, existe o risco de podar a árvore no início de seu desenvolvimento, e potencialmente perder informações valiosas. Crescer a árvore ao seu tamanho máximo e depois podá-la é um processo mais lento, porém gera árvores mais confiáveis [46].

2.5.3 Redes neurais artificiais

O trabalho em redes neurais foi motivado pelo reconhecimento de que o cérebro humano computa informações de forma diferente de um computador convencional [31]. Sua modelagem pode ser feita de forma linear ou não-linear, o que pode ser decisivo se o problema estudado for naturalmente não-linear; sua abordagem é não paramétrica, no sentido em que não é necessário fazer suposições sobre a a distribuição dos dados; redes neurais são adaptáveis e podem ser retreinadas, alterando os pesos sinápticos, o que faz dela uma boa escolha para ambientes não estacionários [53].

Redes neurais artificiais são compostas de múltiplos nós conectados, o que imita a disposição dos neurônios de um cérebro humano. Os nós são conectados e interagem entre si. Esses nós podem receber uma entrada e realizar operações, que em seguida são transmitidas para um próximo nó ou neurônio. A saída de cada nó é feita através do uso de uma função, chamada de função de ativação. Cada conexão entre os nós possui um peso associado, que, quando ajustado, é onde ocorre o "aprendizado" [53].

Redes neurais são modelos estatísticos não-lineares e podem ser usados para regressões ou classificações [45]. Os modelos criados com redes neurais são modelos que operam em dois estágios. Em uma rede neural de uma camada com *back-propagation* ou *single layer perceptron*, e um problema de classificação com K classes, existem K unidades de entrada, sendo que a k-ésima unidade (X_k) modela a probabilidade da classe k; existem ainda K medidas alvo f_k, $k = 1, ..., K$, cada uma sendo codificada como 0 ou 1; características derivadas Z_m são criadas a partir de combinações lineares das unidades de entrada (X_k) e o alvo f_k é modelado como uma função da combinação linear dos Z_m:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \ldots, M \tag{2.11}$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \ldots, K \tag{2.12}$$

$$f_k(X) = g_k(T), k = 1, \ldots, K \tag{2.13}$$

Em que, $Z = (Z_1, Z_2, \ldots, Z_M)$, $T = (T_1, T_2, \ldots, T_K)$ e α_{0m} e β_{0k} são os interceptos do modelo. Usualmente a função de ativação utilizada é o sigmóide [45]:

$$\sigma(\upsilon) = 1/(1 + e^{-\upsilon}) \tag{2.14}$$

A Figura 2.1 exemplifica o caso de uma classificação com K classes (alvo), p entradas e M "neurônios". A camada de entrada, X_p, simboliza a leitura de atributos que contêm informações dos dados analisados. A segunda camada, Z_m, simboliza a camada escondida, em que ocorre o ajuste dos pesos da rede neural. A terceira e última camada, Y_k, simboliza a classificação realizada, baseada nos pesos da camada escondida.

2.5.4 *Naive Bayes*

Em aprendizagem supervisionada, classificadores treinam em observações que contêm rótulos para depois indicar a qual classe uma observação não vista anteriormente pertence. Classificadores também podem ser entendidos como computações de conjuntos de funções discriminantes, um para cada classe e desse conjunto de funções, indica-se que a observação pertence à classe cuja função discriminante é máxima. Seja E uma observação e

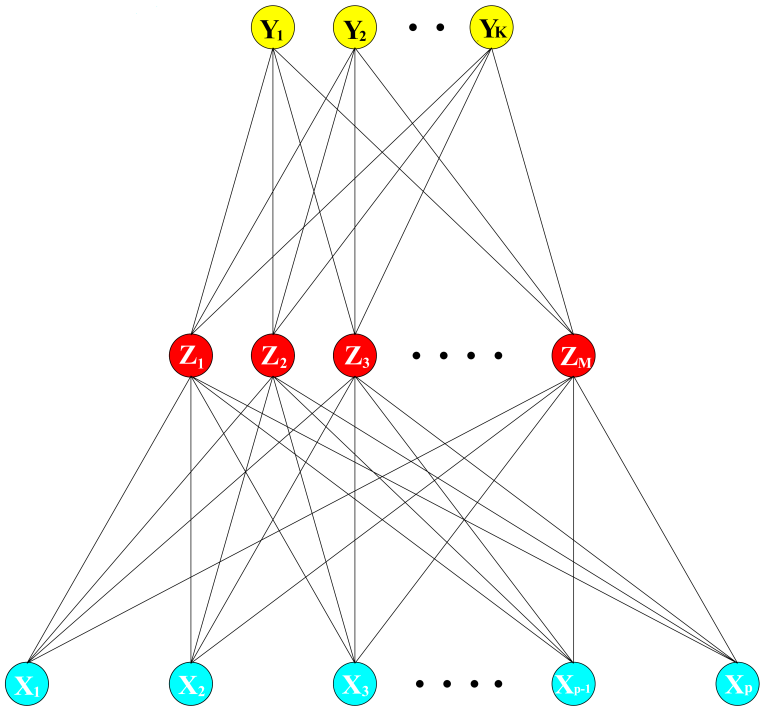

Figura 2.1: Rede neural com uma camada escondida (Fonte: [45]).

$f_i(E)$ a função discriminante correspondente à classe i, a observação será indicada como pertencente à classe C_k, cujo:

$$f_k(E) > f_i(E), \forall i \neq k \tag{2.15}$$

Supondo ainda que a observação é constituída por um vetor de a atributos e υ_{jk} é o valor do atributo A_j na observação, então um possível conjunto de funções discriminantes é dado por:

$$f_i(E) = P(C_i) \prod_{j=1}^{a} P(A_j = \upsilon_{jk}|C_i) \tag{2.16}$$

O classificador obtido através do uso das funções discriminantes da Equação 2.16 é usualmente chamado de *Naive Bayes* [52]. Se os atributos forem independentes, dada uma classe, $P(E|C_i)$ pode ser decomposto como na Equação 2.16, e utilizando o teorema de *Bayes*, tem-se que

$$P(C_i|E) = f_i(E) \tag{2.17}$$

Portanto, $f_i(E)$ retorna a probabilidade do exemplo pertencer à classe C_i. O algoritmo

é chamado de ingênuo, pois ele baseia-se no princípio de que os atributos são independentes dado um classificador, algo que raramente acontece [31].

Nos casos em que os atributos não são independentes entre si, o algoritmo não computa probabilidades, pois sua suposição inicial é quebrada e, portanto $P(C_i|E) \neq f_i(E)$. Nessas situações, dado um exemplo, o algoritmo ainda pode minimizar o erro de classificação. De fato, apesar dessa limitação, *Naive Bayes* obtém um desempenho bom em uma variedade de domínios, incluindo alguns em que existe uma clara dependência entre os atributos de uma classe [52].

Capítulo 3

Plano de trabalho

Historicamente, a modelagem do processo de descoberta de conhecimento começou na academia [41]. Um expoente dos modelos criados na academia é o proposto por Fayyad *et al.* [35], que propôs sua primeira estrutura básica. Não tardou para que viessem modelos propostos por consórcios industriais. Entre eles, dois se destacaram [41], um criado por Cabena *et al.* com suporte da IBM [42] e o modelo industrial de seis etapas, CRISP-DM [43], desenvolvido por um grande consórcio de companhias europeias. O CRISP-DM acabou tornando-se o modelo mais utilizado para o processo de mineração de dados [41].

Neste trabalho será utilizado o modelo CRISP-DM, definido com seis etapas como pode ser visto na Figura 3.1 e detalhado em [43, 36]. Resumidamente, para este trabalho, o processo implica em entender o negócio para poder entender os dados, em seguida organizar os dados de forma que seja possível classificar as observações em evasão e não evasão. Finalizando o processo, avalia-se quais classificadores se comportam melhor neste trabalho e um *framework* é proposto para que o processo possa ser repetido de forma simplificada. Em conjunto com a descrição das etapas do processo, serão descritos os esforços realizados no trabalho.

3.1 Entendimento do negócio

Esta etapa foca em entender os objetivos e requerimentos de uma perspectiva de negócios. Deve-se também converter esse entendimento em uma definição de problema de mineração de dados e desenvolver um planejamento para atingir os objetivos. Ela é dividida em:

- determinar os objetivos de negócio;
- determinar os objetivos da mineração de dados e
- gerar um planejamento preliminar para atingir os objetivos.

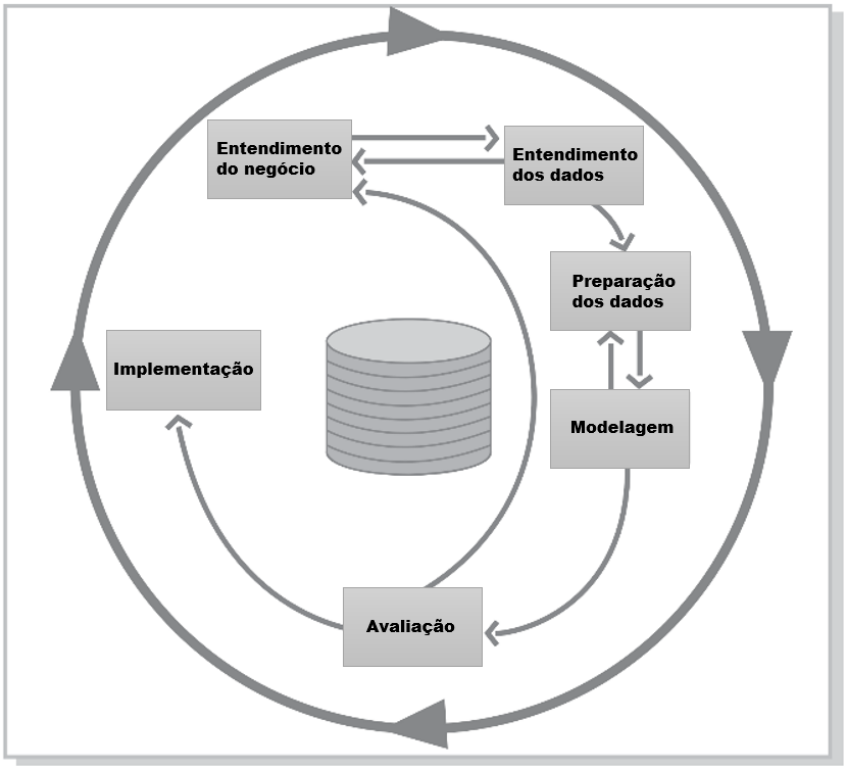

Figura 3.1: O modelo de processos CRISP-DM (Fonte: [43]).

Grande parte do entendimento de negócio está contido nos Capítulos 1 e 2 deste trabalho. O Capítulo 1 introduz a problemática da evasão com foco no ensino superior, define os objetivos do trabalho e justifica sua necessidade. Já o Capítulo 2 traz uma revisão de literatura, resumindo alguns trabalhos de mineração de dados focados em evasão no ensino superior e descreve os algoritmos de mineração de dados utilizados neste trabalho.

O objetivo geral deste trabalho é identificar as características de alunos que evadem do ensino superior brasileiro, mais especificamente, os de graduação presencial. Para atingir esse objetivo, primeiramente se faz necessário definir quais alunos serão rotulados como "evasão" na base de dados. A definição utilizada no trabalho é melhor detalhada na etapa de "Preparação dos dados" (Seção 3.3), pois é necessário que o leitor tenha um maior entendimento dos dados do CES. Nesta etapa, serão explicados os três níveis de evasão que serão utilizados:

- evasão a nível da IES;
- evasão a nível do curso e
- evasão a nível da área de estudo.

A área de estudo utilizada neste trabalho é a definida na *International Standard Classification of Education* (ISCED) adaptada para cursos do Brasil [54]. Ela utiliza um

Tabela 3.1: Áreas gerais de estudo.

Código	Área Geral
0	Programas ou cursos gerais
1	Educação
2	Humanidades e Artes
3	Ciências Sociais, Negócios e Direito
4	Ciências, Matemática e Computação
5	Engenharia, Produção e Construção
6	Agricultura e Veterinária
7	Saúde e Bem Estar Social
8	Serviços

Tabela 3.2: Exemplos de evasão em diferentes níveis.

Aluno	Curso	Área de estudo	Evasão	IES	Nível de evasão
João	Matemática	4	Sim	UnB	Curso
João	Física	4	Não	UnB	–
Ricardo	Engenharia	5	Sim	UnB	Curso e Área
Ricardo	Direito	3	Não	UnB	–
Ana	Medicina	7	Sim	UnB	Curso, IES e Área

código de três dígitos em um sistema hierárquico para classificar as áreas de formação. O primeiro dígito indica a grande área ou área geral, o segundo dígito indica a área específica ou simplesmente "área", o terceiro dígito indica a área detalhada ou sub-área. Há 9 áreas gerais, 25 específicas e 80 áreas detalhadas. Serão utilizadas as 9 áreas gerais de estudo, conforme a Tabela 3.1.

A evasão a nível da IES se dá quando um aluno deixa de possuir pelo menos um vínculo com qualquer curso daquela instituição. Já a evasão a nível do curso se dá quando um aluno evade de um determinado curso, podendo por exemplo manter-se vinculado à mesma IES em outro curso. Finalmente, a evasão a nível de área de estudo acontece quando um aluno deixa de possuir vínculo com pelo menos um curso de determinada área de estudo.

Na Tabela 3.2 podem ser vistos exemplos dos níveis de evasão, sendo que cada linha é chamada de vínculo do aluno ao curso. O aluno João, por exemplo, possui dois vínculos, um com o curso de Matemática e outro com o curso de Física. No exemplo o aluno pode ter se transferido entre os cursos, o que explicaria ele possuir os dois vínculos. Como a evasão ocorre apenas no curso de Matemática, pois o curso de Física está na mesma área de estudo e IES, ela é classificada como evasão a nível de curso. No caso do aluno Ricardo, há uma evasão a nível de curso e área de estudo, pois como o João, ele se transfere dentro da mesma instituição, mas deixando sua área de estudo original. O último exemplo mostra a aluna Ana, que possui apenas um vínculo com a IES. Este vínculo é um de evasão e como não há outros, a aluna é classificada como evasão a nível de curso, área de estudo

e IES.

Além da definição do nível de evasão, o planejamento inicial do projeto de mineração inclui a seleção de um grupo de dados em que os classificadores possam ser treinados. Inicialmente, são selecionados os alunos da UnB da massa de dados do CES. A escolha da UnB se dá por alguns motivos a saber. Trata-se de uma unidade de observação de um universo de instituições de ensino superior que possuem informações em todos os anos coletados, em que o processo de mineração de dados pode ser testado e avaliado adequadamente. Também por ser uma universidade federal, ou seja, gerida com dinheiro público e por ser a universidade em que este trabalho é realizado.

Dos alunos da UnB, seleciona-se apenas os ingressantes de cada ano disponível nas bases do Inep, reduzindo a base trabalhada. A utilização dos alunos ingressantes se justifica por vários autores [20, 24, 28, 29, 55, 56, 57, 58] identificarem em seus trabalhos que o primeiro ano após o ingresso é o ano de maior risco de evasão.

O próximo passo é unir às informações do CES os atributos do ENEM. Os alunos que possuem informações em ambas as bases são selecionados. Após a conclusão da construção das *tabelas* de dados finais (uma para cada nível de evasão), serão treinados cinco tipos de classificadores, C5.0, *Naive Bayes*, redes neurais, regressão logística e CART. A arquitetura desse *framework* pode ser vista na Figura 3.2. Os detalhes da construção das *tabelas* e do treinamento dos classificadores são detalhados nas etapas subsequentes.

Reforça-se que, como definido no Capítulo 1, o objetivo geral é identificar as principais características de alunos que evadem do ensino superior brasileiro através do uso de técnicas de mineração de dados. Também foi gerado um planejamento preliminar para atingir esses objetivos, com a escolha das observações que serão utilizadas e a definição dos níveis de evasão que serão analisados.

3.2 Entendimento dos dados

A etapa começa com uma coleta inicial de dados e subsequente familiarização. A identificação de problemas de qualidade dos dados também é realizada nesta etapa. Ela é dividida em:

- coleta de dados inicial;
- descrição dos dados;
- exploração dos dados e
- verificação da qualidade dos dados.

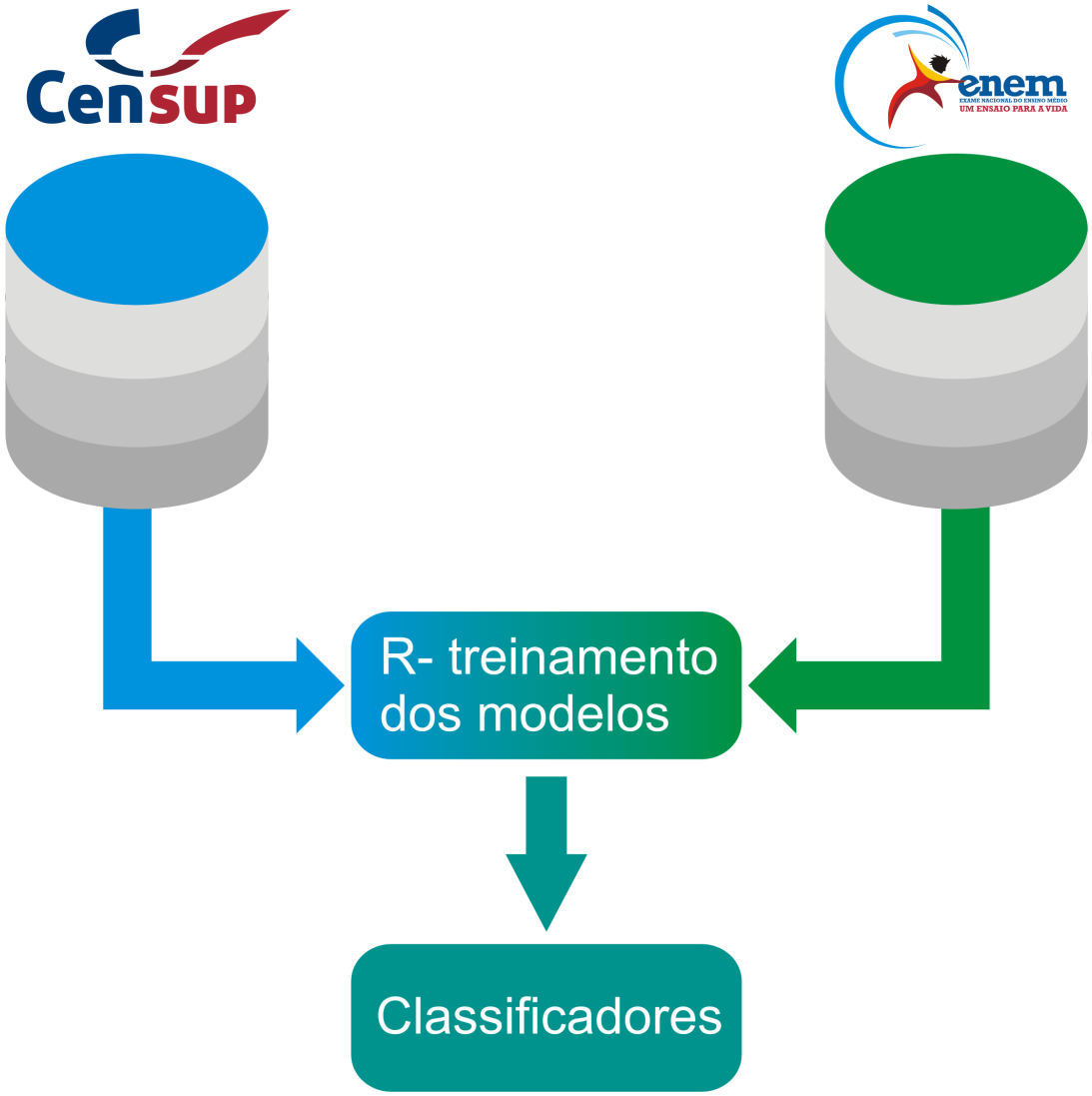

Figura 3.2: Diagrama da arquitetura do *framework* proposto.

O Censo da Educação Superior é realizado anualmente e tem como referência o decreto n^o 6.425 [9], que obriga as instituições a prestarem informações de seus alunos, cursos, docentes e instalações. As IES devem informar seus dados referentes ao ano letivo anterior ao ano de coleta, ou seja, o censo de 2014 é coletado em 2015. A base de dados do CES contém todos os alunos e docentes do ensino superior brasileiro identificados por CPF, as únicas exceções são os alunos estrangeiros, que não tem obrigatoriedade de possuir o cadastro de pessoa física [59]. Algumas das informações das IES e dos cursos não são preenchidas pelas instituições durante o período do censo. Informações como o código do curso, sua área de estudo, seu grau acadêmico (bacharelado, licenciatura e tecnológico), sua modalidade de ensino (presencial e à distância), entre outras fazem parte de um

cadastro mantido pelo MEC, intitulado de e-MEC[1]. Esse cadastro é reflexo das regulações feitas pelo MEC no ensino superior. Um curso novo, por exemplo, só aparecerá no censo, se antes ele tiver sido autorizado a funcionar e tiver sido incluído no cadastro e-MEC [60].

Ao consolidar as informações coletadas ao longo do ano, o Inep disponibiliza anualmente em seu *site* cinco arquivos de dados[2], são eles: "alunos", "cursos", "IES", "docentes" e "locais de oferta", dos quais três são utilizados neste estudo, "alunos", "cursos" e "docentes". Os arquivos publicados pelo Inep são arquivos de texto extraídos a partir de *tabelas SQL* e não estão em forma normal como definida em [31]. Dessa forma, são necessárias menos *tabelas* para obter as informações relativas aos alunos. O código da IES, por exemplo, se encontra em todas as *tabelas*; a categoria administrativa da IES se encontra nas *tabelas* de "alunos", "cursos" e "IES" retirando a necessidade de se utilizar a *tabela* de "IES" para obter esses atributos. Outro exemplo é o município em que o curso é oferecido, que se encontra na *tabela* de "cursos", não sendo necessário o uso da *tabela* de "locais de oferta". As *tabelas* utilizadas neste trabalho, são as mesmas disponibilizadas pelo Inep em seu *site*, exceto pela presença do CPF que identifica alunos e docentes e não está presente nos arquivos públicos.

As *tabelas* de alunos do CES de 2010 a 2014 possuem milhões de observações por ano. Cada observação nessa *tabela* é referente a um par aluno/curso, chamado de vínculo do aluno ao curso [59]. As IES informam ao censo apenas a última informação de cada aluno em cada curso, dessa forma cada estudante possui no máximo um vínculo com cada curso. Por exemplo, um aluno que ingressou em Matemática mas deixou o curso no mesmo ano letivo é representado na *tabela* por apenas uma observação. Além disso, nada impede que um aluno possua vínculos a vários cursos. No caso de um aluno que saiu de um curso de engenharia e, no mesmo ano, foi cursar física, por exemplo, deverão constar duas observações na *tabela* de aluno, uma referente ao curso de engenharia e outra referente ao curso de física. Para cada vínculo de aluno ao curso existe um atributo que indica a situação desse aluno no curso. É através da situação do aluno no curso que se faz possível, neste trabalho, determinar quais alunos são considerados evasão e quais não. Na Tabela 3.3, pode-se ver como a informação é preenchida pelas IES de acordo com algumas situações hipotéticas.

De acordo com o manual de preenchimento das informações de aluno no CES [59], as possíveis situações de vínculo do aluno com um curso são:

- cursando: situação de vínculo do aluno que não concluiu a totalidade da carga horária exigida para a conclusão do curso, no ano de referência do Censo.

[1] `http://emec.mec.gov.br/`
[2] `http://inep.gov.br/microdados`

Tabela 3.3: Exemplos de preenchimento do censo baseado em uma IES com dois semestres.

Curso	Aluno	1º semestre	2º semestre	Situação	Tipo de vínculo
Administração	Manuel	Cursando	Trancado	Trancado	Ativo
Administração	Viviane	Trancado	Cursando	Cursando	Ativo
Administração	Paula	Formado	-	Formado	Final
Administração	Douglas	Cursando	Cursando	Cursando	Ativo
Administração	Luciana	Transferido	-	Transferido	Final
Direito	Luciana	-	Cursando	Cursando	Ativo

- matrícula trancada: aluno que, no ano de referência do Censo, está com a matrícula trancada na IES.
- desvinculado: aluno que, na data de referência do Censo, não possui vínculo com o curso por motivos de abandono, desligamento ou transferência para outra IES.
- transferido para outro curso da mesma IES: aluno que foi transferido para outro curso de graduação da mesma IES. Se o aluno possuir esta situação, ele necessariamente deve possuir outro vínculo na mesma IES com situação diferente de transferido.
- formado: aluno que concluiu a totalidade dos créditos acadêmicos exigidos para titulação no curso durante o ano de referência do Censo. Não é obrigatório que o aluno tenha realizado a colação de grau e/ou participado do Exame Nacional de Desempenho de Estudantes (ENADE).
- falecido: aluno falecido durante o ano de referência do Censo.

Na Tabela 3.4, pode ser vista a quantidade de observações que existem nas *tabelas SQL* de aluno do CES agrupadas por situação de vínculos dos alunos aos cursos nos anos de 2010 a 2014. Pode-se notar que a quantidade de vínculos com a situação de "formado" não acompanha o crescimento de vínculos na *tabela* ao longo dos anos. Enquanto o total cresce cerca de 600.000 vínculos por ano, a quantidade de "formados" chega a cair em 2013 para voltar a um valor próximo do valor de 2012 no último ano analisado. A maior parte dos vínculos encontram-se em "cursando" e "matrícula trancada", chamados vínculos ativos. Os demais vínculos são chamados de vínculos finais, pois não se espera que o aluno apareça no mesmo curso depois de apresentar uma situação de "desvinculado", "transferido", "formado" ou "falecido".

Além da situação de vínculo do aluno ao curso existem outras informações à respeito de suas vidas acadêmicas na *tabela* "alunos". Por exemplo, o semestre e ano em que o aluno ingressou no curso, a forma como ele ingressou, ou seja, por vestibular, por programa de avaliação seriada, através de vagas remanescentes entre outras formas. Se é um aluno

Tabela 3.4: Número de observações por situação de vínculo nas tabelas de aluno do CES de 2010 a 2014.

Situação \Ano	2010	2011	2012	2013	2014
Cursando	5.427.071	5.742.829	6.002.015	6.328.152	6.809.245
Matrícula trancada	683.761	759.512	971.097	1.039.665	1.211.342
Desvinculado	1.138.298	1.332.298	1.440.755	1.469.176	1.632.828
Transferido	106.120	103.022	94.379	96.182	108.796
Formado	980.662	1.022.711	1.056.069	994.812	1.030.520
Falecido	1.307	1.352	1.168	1.302	1.204
Total	**8.337.219**	**8.961.724**	**9.565.483**	**9.929.289**	**10.793.935**

que ingressou através de reserva de vagas, ou seja, cotas por etnia, deficiência ou por estudar em escola pública no ensino médio. Existem também atributos que indicam se os alunos fazem atividades extracurriculares como monitoria, estágio e pesquisa. No total, as quantidades de atributos das *tabelas* são de 90, 98, 106, 114 e 122 para os anos de 2010 a 2014, nesta ordem.

As outras *tabelas* ("docentes", "cursos", "IES", "locais de oferta") possuem atributos sobre as características dos cursos, dos docentes e das IES a que esses alunos estão vinculados. Para as IES, existem informações como localização (UF e município), quantidade de funcionários técnicos, categoria administrativa (pública, privada), organização acadêmica (universidade, faculdade entre outras) e informações sobre as bibliotecas que atendem aos alunos. Para os cursos, alguns dos atributos são o grau acadêmico do curso (bacharelado, licenciatura ou tecnológico), modalidade de ensino (presencial ou a distância), um indicador para cursos gratuitos, a área de estudo do curso ou ISCED adaptada [54]. Para os docentes, existem informações de escolaridade; idade; sexo; cor/raça; nacionalidade; regime de trabalho, que pode ser tempo integral com dedicação exclusiva, sem dedicação exclusiva, tempo parcial ou horista; entre outras.

As *tabelas* do CES possuem informações relativas a vida do estudante na instituição, porém poucos atributos são relativos ao indivíduo. Os atributos que existem são cor/raça, sexo, idade, nacionalidade e deficiências físicas. Para que seja possível identificar um perfil mais detalhado dos alunos evadidos é necessário obter mais informações sobre os estudantes. Por essa razão, a base de dados formada através do Exame Nacional do Ensino Médio (ENEM) também é utilizada no trabalho.

O ENEM é um exame de âmbito nacional, criado em 1998, com o intuito de avaliar a qualidade do ensino médio brasileiro [61]. Em 2009, um novo modelo de exame foi implementado, buscando a unificação dos processos seletivos das universidades federais brasileiras [62]. No mesmo ano, foi implementado o Sistema de Seleção Unificada (SISU) [63], que é a plataforma pela qual os estudantes se candidatam a estas universidades. O ENEM também é utilizado para aquisição de bolsa de estudo integral ou parcial

em instituições privadas através de programas como Programa Universidade para Todos (ProUni) e Fundo de Financiamento ao Estudante do Ensino Superior (FIES) [64, 65].

Ao realizar a inscrição para o exame, os estudantes devem preencher um questionário de até 76 perguntas, dividido em duas partes. Na primeira, todos os inscritos devem responder, enquanto que na segunda parte, apenas os estudantes que vão requerer a certificação do ensino médio respondem. Nos anos imediatamente seguintes aos exames, o Inep consolida as bases de dados que contêm as respostas dos estudantes aos questionários, suas notas nas quatro áreas do exame (ciências da natureza; ciências humanas; linguagens e códigos; matemática) e a nota da redação. É importante ressaltar que, de maneira geral, qualquer pessoa pode fazer o exame. Em casos específicos, como o de detentos, também é possível realizar a prova, com a ressalva de que é necessário comunicar a situação no momento da inscrição. Estudantes de ensino médio que não irão se formar no mesmo ano também podem fazer o exame com a ressalva de serem treineiros e não poderem utilizar os resultados para ingressar no ensino superior [61, 62, 63].

Combinando os dados do ENEM com os do CES, obtém-se as características do estudante antes de ingressar no ensino superior, assim como algumas informações das vidas acadêmicas dos alunos nas IES. Vale notar que não se tem notícia, até a data deste trabalho, de estudos sobre evasão no ensino superior que tenham utilizado essas fontes de dados.

As bases do ENEM possuem apenas uma *tabela* de interesse por ano e assim como as *tabelas* de aluno do CES, possuem milhões de observações nos anos de 2010 a 2014. Cada observação na *tabela* do ENEM corresponde a um inscrito no exame. Na Tabela 3.5 pode-se ver a quantidade de inscritos por ano e quantos desses inscritos possuem notas em todas as quatro provas. Os inscritos que não possuem informação nas quatro áreas são aqueles que não compareceram nos dias de prova. Além das notas nas áreas do exame, os questionários trazem informações sócio econômicas dos inscritos, como escolaridades do pai e da mãe; se exerce ou não atividade remunerada; a renda de sua família entre outras. Os questionários sofreram diversas mudanças ao longo dos anos estudados.

O questionário de 2011 possui um total de 57 questões, sendo 33 referentes a primeira parte, obrigatória para todos inscritos, e 24 referentes a segunda parte, preenchida pelos inscritos que requereram a certificação do ensino médio. Em 2011, o questionário sofreu modificações, tendo 75 questões no total. A primeira parte manteve as mesmas 33 questões de 2010 e à segunda parte foram adicionadas questões, totalizando 42 questões. Em 2012 houve nova alteração, o questionário passou a possuir 62 questões no total, 39 referentes a primeira parte e 23 à segunda. No ano de 2013, ocorreu a última mudança. O questionário ficou com 76 perguntas no total, sendo 53 referentes a primeira parte e 23 referentes a segunda.

Tabela 3.5: Número de observações e quantidade de inscritos com nota nas 4 áreas nas bases do ENEM de 2010 a 2014.

Situação \Ano	2010	2011	2012	2013	2014
Total de observações	4.626.094	5.366.948	5.791.065	7.173.584	8.422.248
Inscritos com notas	3.242.776	3.853.330	4.079.886	5.007.953	5.947.914

A quantidade de atributos nas bases de dados do ENEM são 148, 165, 154, 169 e 168, respectivamente para os anos de 2010 a 2014. Além das notas nas quatro áreas do exame e das respostas aos questionários, existem também atributos de identificação do aluno e informações como município de residência, município onde o inscrito frequenta escola, informação de deficiência do inscrito entre outros. As mudanças ocorridas nos questionários do ENEM implicam que as informações de diferentes anos não sejam imediatamente compatíveis para o uso de treinamento de classificadores. Portanto, se faz necessário realizar uma reorganização das informações do questionário, que é detalhada na seção de preparação dos dados. A lista completa dos atributos pode ser encontrada na página do projeto deste trabalho[3].

A coleta dos dados que foram utilizados neste trabalho, base do Exame Nacional do Ensino Médio com identificação por CPF e base do CES também identificados por CPF para os anos de 2010 a 2014 foi realizada junto ao Inep. A maioria dos atributos são qualitativos nas duas bases. Na base do CES, apenas a data de ingresso do aluno no curso, a quantidade de vagas e o número de candidatos não são qualitativos. Na base do ENEM, apenas as notas dos alunos nas quatro áreas do conhecimento, a nota da redação, as cinco notas de competências de escrita e a idade do aluno são atributos quantitativos.

Dados identificados dos alunos do ensino superior existem de 2009 a 2014, última coleta divulgada. A base de dados de 2009 não é utilizada para este trabalho, pois ela possui alguns problemas de consistência. Existem alunos e docentes sem identificação de CPF, o que compromete a ligação dessa *tabela* com a do ENEM. Além disso, a não obrigatoriedade de CPF permitiu a criação de registros duplicados de alunos e docentes. Um outro problema é que a integração entre o sistema do censo e o e-MEC não funcionou como deveria. Para finalizar a coleta dentro do prazo estabelecido à época, foi necessário permitir que as próprias IES cadastrassem seus cursos. Esse tipo de prática acarreta na criação equivocada de cursos e por consequência, de informações equivocadas sobre alunos e docentes. Como a coleta de 2009 foi a primeira realizada dessa forma (individualizada e não agregada), existem outros detalhes que foram modificados nos anos posteriores para adequar melhor a coleta à realidade das IES, como a informação do turno em que o aluno cursa, que não existia em 2009 e a remoção da situação de "provável formando" que em 2009 causava ambiguidade no momento de preenchimento da situação do aluno.

[3] `https://github.com/lucke71/Classify_dropout/tree/master/Dicionarios%20ENEM`

Em relação à verificação da qualidade dos dados, toda coleta do CES é feita em etapas, sendo a etapa final a de verificação da qualidade da informação prestada pelas IES. Além disso, o sistema de coleta do CES possui diversas regras de sistema que previnem inconsistências das informações. Por exemplo, o sistema não permite que um curso seja informado com alunos cursando e não exista pelo menos um docente em atividade vinculado a este curso. Na base do ENEM o processo é similar, uma vez que a equipe do Inep procura inconsistências na base de dados, como por exemplo, idades como 130 anos ou menos de 10 anos. Em relação às respostas aos questionários, não há muito o que se fazer em termos de qualidade dos dados, pois a resposta de cada pessoa aos itens não é passível de verificação.

Algumas inconsistências foram encontradas na *tabela* do ENEM relativa ao ano de 2010. Algumas informações que deveriam possuir valores estavam em branco. Um exemplo é a questão 02 do questionário socioeconômico, que pergunta para todos os inscritos a escolaridade do pai. Essa questão é de resposta obrigatória e foi feita para todos os inscritos, no entanto, o atributo possui observações sem valor preenchido. Na próxima seção, será detalhada a quantidade de observações que possuem dados faltantes, assim como o tratamento realizado nessas observações.

Nesta etapa, foi explicada a estrutura dos dados do ENEM e do CES, assim como parte da coleta dessas informações. Além disso, uma exploração de dados foi conduzida, mostrando dados descritivos de ambas as bases. Também foi realizada uma verificação da qualidade dos dados, em que se percebeu que as *tabelas* do ENEM possuíam alguns problemas, seja por inconsistência ou por características da coleta de informações. Não foram encontradas inconsistências nas *tabelas* do CES.

3.3 Preparação dos dados

Esta etapa cobre todos os passos necessários para construir a *tabela* de dados final, onde serão executados os métodos de mineração de dados e aprendizado de máquina. Ela é dividida em:

- seleção dos dados;
- limpeza dos dados;
- construção dos dados;
- integração dos dados e
- formatação das subetapas dos dados.

Inicialmente, de acordo com o planejamento inicial, buscou-se organizar apenas uma *tabela* com os dados dos alunos ingressantes da Universidade de Brasília (UnB) de cursos de graduação presencial de 2010 a 2014, para posteriormente formar *tabelas* para cada um dos níveis de evasão. Ou seja, primeiramente trabalhou-se apenas com as informações do CES, criando o grupo alvo do estudo para em um segundo momento incorporar os atributos do ENEM.

Todas as *tabelas* são acessadas e manipuladas através de um servidor SQL Oracle. Em seguida, elas são levadas para o ambiente do R, em que é feito o treinamento dos classificadores. Os dados manipulados no servidor SQL são levados para o *software* R através de uma conexão entre os dois. Existe mais de uma forma de realizar essa conexão, neste trabalho, foi utilizado o pacote *RODBC*[4]. O acesso aos dados é feito dentro das instalações físicas do Inep, pois os dados possuem identificação dos estudantes o que torna a informação sigilosa. O acesso para pessoas externas à autarquia é feito através de uma sala segura onde os dados são disponibilizados. Para mais detalhes sobre como ter acesso aos dados, verificar o *site* do Inep[5].

Para explicar a construção da *tabela* agregada com dados do CES de 2010 a 2014 o texto foi dividido em duas partes, seleção de atributos e seleção de observações, no entanto, no *script* SQL, não houve necessidade de fazer essa separação. As observações e atributos são escolhidos em um mesmo comando. Além disso, os procedimentos de seleções de atributos e observações são feitos de forma que são repetidos ano a ano para 2010 a 2014, observando algumas diferenças nas *tabelas* de cada ano. O nível de detalhe do texto a seguir é alto, pois um dos objetivos é produzir um *framework* do trabalho realizado, que possa ser utilizado por outros pesquisadores.

Primeiramente, foram selecionados os atributos que compõem a *tabela* agregada. Neste passo, foi necessário utilizar três *tabelas* do CES, "alunos", "cursos" e "docentes". A partir da *tabela* de "alunos", são selecionados vinte e nove atributos. Quatro são atributos de identificação: código do aluno na base do CES, CPF, código da IES, código do curso ao qual o aluno possui vínculo e um atributo criado que serve para identificar o ano relativo às informações. O restante dos atributos são de características dos alunos ou de seus cursos. Os atributos relacionados aos alunos são: ano e semestre de ingresso, ambos extraídos da data de ingresso do aluno no curso; nacionalidade do aluno; sua situação no curso; indicador de ingresso por reserva de vagas, como cotas étnicas ou de deficiência; quatro indicadores de atividades extracurriculares, estágio, monitoria, pesquisa e extensão; quatro indicadores de bolsa relacionados às quatro atividades extracurriculares; número de cursos do aluno dentro da UnB e o número de IES a que o aluno está vinculado em toda a

[4] `https://cran.r-project.org/web/packages/RODBC/index.html`

[5] `http://inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/pesquisadores-ja-tem-acesso-a-informacoes-protegidas-do-inep/21206`

base do CES no ano em questão, sendo as duas últimas construídas a partir de contagem dos vínculos nas *tabelas* de cada ano. Os atributos relacionados aos cursos são: área geral de estudo do curso; quatro indicadores de turno de funcionamento dos cursos (matutino, vespertino, noturno e integral), assim como quatro prazos previstos para que os alunos terminem sua graduação (tempo de integralização do curso) e um indicador de curso de Área Básica de Ingresso (ABI).

Os cursos do tipo ABI agregam ingressantes para que em um segundo momento os alunos possam escolher uma formação acadêmica específica. Um exemplo de curso ABI pode ser visto nos cursos de Letras, em que todos alunos ingressam em um único curso básico para depois decidirem em qual linguagem se especializarão e se formarão [59]. Além de servir como entrada única para algumas formações acadêmicas, cursos ABI possuem características distintas dos demais cursos. Não há formados em cursos ABI, pois os alunos não recebem um diploma por finalizar o ciclo básico de disciplinas [59]. Dessa forma, quando um aluno segue o fluxo normal a partir de um curso ABI é necessário que ele passe a cursar em uma formação acadêmica de sua escolha para que possa se formar.

Nos dados do CES, essa passagem de um curso ABI para um curso de formação se dá através de uma transferência interna, ou seja, o vínculo do aluno assume a situação de "transferido para outro curso da mesma IES" no curso ABI e a situação de "cursando" em seu curso de destino. Também é importante notar que os cursos ABI foram classificados em 2010, como pertencentes a apenas uma área geral de estudo, "Programas ou cursos gerais". No entanto, nos anos seguintes, sua classificação nas *tabelas* do CES não seguiu o mesmo comportamento. Alguns cursos ficaram com a classificação em branco enquanto outros entraram em alguma outra área, pois a área geral de estudo "Programas ou cursos gerais" deixou de ser utilizada nos anos seguintes a 2010.

A partir da *tabela* de "cursos", são selecionados quatro atributos: número de vagas oferecidas, construído a partir da quantidade de vagas oferecidas em cada turno; número de candidatos ao curso, também construído a partir da quantidade existente em cada turno; categoria administrativa e organização acadêmica da IES. No caso de o grupo de alunos selecionados for de apenas uma instituição, a categoria administrativa e a organização acadêmica não produzem ganho de informação para os classificadores, pois eles são iguais para todas observações. Esses dois atributos são incluídos para os casos em que são selecionados alunos de mais de uma IES.

Da *tabela* de "docentes", são criados 17 atributos, todos contagens das quantidades de professores por suas características, agrupados por cursos. Dessa forma, informações relativas aos docentes de cada curso são agregadas e, portanto, podem ajudar a descrever o curso a que o aluno está vinculado.

Cinco atributos de contagem são formados a partir das cinco possíveis situações de um docente em um curso:

1. em exercício;
2. afastado para capacitação;
3. afastado para exercício em outros órgãos/entidades;
4. afastado por outros motivos e
5. afastado para tratamento de saúde.

São criados mais cinco atributos de contagem agrupados por cursos a partir da escolaridade dos docentes:

1. sem graduação;
2. graduado;
3. com especialização;
4. mestre e
5. doutor.

Outros quatro atributos de contagem são criados a partir do regime de trabalho dos docentes:

1. integral com dedicação exclusiva;
2. integral sem dedicação exclusiva;
3. tempo parcial e
4. horista.

Os últimos três atributos de contagem são criados a partir da nacionalidade do docente, que possuem os mesmos valores na tabela de alunos:

1. brasileiro;
2. brasileiro naturalizado e
3. estrangeiro.

Tabela 3.6: Número de vínculos de alunos por situação e área geral de estudo do curso na UnB de 2010 a 2014.

	Situação de vínculo						
Área geral	Cursando	Trancado	Desvinculado	Transferido	Formado	Falecido	**Total**
–	1.756	120	298	306	19	0	**2.499**
0	636	16	110	360	0	2	**1.124**
1	29.873	2.746	7.149	2.076	5.591	8	**47.443**
2	7.689	766	1.410	307	1.311	3	**11.486**
3	35.260	2.492	4.330	323	6.138	8	**48.551**
4	9.772	858	1.918	151	1.543	4	**14.246**
5	20.651	1.879	2.384	1.318	1.635	2	**27.869**
6	5.609	402	615	21	763	1	**7.411**
7	15.957	1.023	1.553	66	1.775	1	**20.375**
8	1.718	134	226	6	96	1	**2.181**
Total	**128.921**	**10.436**	**19.993**	**4.934**	**18.871**	**30**	**183.185**

Em seguida, a seleção das observações é realizada. Utiliza-se a *tabela* de "alunos" do CES para selecionar apenas os alunos que possuíam algum vínculo a cursos da UnB. Para fazer isso, basta filtrar pelo código da instituição desejada, um dos atributos da *tabela*. Também é necessário escolher apenas os cursos de graduação presencial, o que é alcançado com mais dois filtros na mesma *tabela* de alunos, um para a modalidade de ensino (presencial ou à distância) e outro para o nível acadêmico (graduação ou sequencial de formação específica)[6]. Um dos cursos da UnB possui mais de uma área de estudo ao longo do período analisado. O curso de gestão do agronegócio foi classificado como sendo da grande área "Agricultura e Veterinária" em 2010 e nos outros anos foi classificado na área "Ciências Sociais, Negócios e Direito". Com o objetivo de evitar problemas nos classificadores de evasão a nível de área de estudo, alterou-se a área de estudo geral do curso em questão para a mesma que ele possui na maior parte dos anos analisados.

Na Tabela 3.6 pode-se ver por situação do aluno e área geral de estudo a quantidade de vínculos que existem na *tabela* criada a partir da agregação dos dados do CES para a UnB de 2010 a 2014. Percebe-se que a UnB possui a maior parte dos seus vínculos em situação de "cursando" e que a área geral de estudos "Ciências Sociais, Negócios e Direito" é a que possui a maior quantidade de vínculos entre as áreas. Na Tabela 3.7 pode-se ver a mesma informação, desta vez apenas com cursos de graduação presencial. Percebe-se que a área geral de estudos "Educação" é a que apresentou a maior queda na quantidade de vínculos se comparada com a Tabela 3.6, o que indica que a área de educação é a que mais oferece cursos à distância dentro da universidade.

[6] A UnB não possui cursos sequenciais de formação acadêmica.

Tabela 3.7: Número de vínculos de alunos por situação e área geral de estudo nos cursos de graduação presencial da UnB de 2010 a 2014.

	Situação de vínculo						
Área geral	Cursando	Trancado	Desvinculado	Transferido	Formado	Falecido	**Total**
–	1.756	120	298	306	19	0	**2.499**
0	636	16	110	360	0	2	**1.124**
1	24.523	2.386	5.415	2.076	4.412	3	**38.815**
2	7.689	766	1.410	307	1.311	3	**11.486**
3	34.634	2.472	3.970	323	5.756	8	**47.163**
4	9.772	858	1.918	151	1.543	4	**14.246**
5	20.651	1.879	2.384	1.318	1.635	2	**27.869**
6	5.609	402	615	21	763	1	**7.411**
7	15.957	1.023	1.553	66	1.775	1	**20.375**
8	1.718	134	226	6	96	1	**2.181**
Total	**122.945**	**10.056**	**17.899**	**4.934**	**17.310**	**25**	**173.169**

O próximo passo para se criar a *tabela* agregada de informações do CES é selecionar os alunos ingressantes de cada ano. O ano de ingresso de cada aluno é extraído da informação de data de ingresso no curso, em seguida o ano de ingresso é igualado ao ano de referência do censo, ou seja, para a *tabela* de 2010, são escolhidos os alunos com ano de ingresso igual a 2010, para *tabela* de 2011, são escolhidos os alunos com ano de ingresso igual a 2011 e assim por diante. Na Tabela 3.8, pode-se ver a quantidade de vínculos de alunos ingressantes de cursos de graduação presencial na UnB por situação do aluno e área geral de estudo. Percebe-se que a situação de vínculo mais comum continua sendo a de "cursando", no entanto sua proporção em relação as outras situações é consideravelmente maior quando comparada a Tabela 3.6 e a Tabela 3.7.

Na Tabela 3.6, o percentual de vínculos com situação de cursando era de 70,38%, na Tabela 3.7 passou a ser de 71%, enquanto que na Tabela 3.8 é de 88,62%. As situações de vínculo que mais caíram percentualmente foram "formado" e "desvinculado", em 8,67% e 4,19%, respectivamente. Em relação às áreas gerais de estudo, a área "Ciências Sociais, Negócios e Direito", continua sendo a área com a maior quantidade de vínculos, sendo que não houve uma grande diferença em termos percentuais em relação à Tabela 3.6 e à Tabela 3.7.

Com os atributos e observações selecionados nas *tabelas* do CES ano a ano, é feita a junção[7] das novas *tabelas* que possuem os atributos de cursos, alunos e docentes, formando cinco *tabelas*, uma para cada ano. A junção das *tabelas* é feita através do código do curso, presente em todas as *tabelas*. Vale ressaltar que os códigos de curso são únicos, portanto

[7]O equivalente a um *left join* da linguagem SQL, ou seja, não se altera a quantidade de observações, apenas a quantidade de atributos.

Tabela 3.8: Número de vínculos de alunos ingressantes por situação de vínculo e área geral de estudo nos cursos de graduação presencial da UnB de 2010 a 2014.

	Situação de vínculo						
Área geral	Cursando	Trancado	Desvinculado	Transferido	Formado	Falecido	**Total**
–	1.446	55	126	3	0	0	**1.630**
0	448	5	27	0	0	0	**480**
1	8.339	492	689	22	184	0	**9.726**
2	2.439	130	206	5	62	0	**2.842**
3	10.442	453	638	1	176	2	**11.712**
4	3.331	130	219	0	168	0	**3.848**
5	6.533	202	366	1	3	0	**7.105**
6	1.531	50	128	0	2	0	**1.711**
7	4.863	198	322	1	2	0	**5.386**
8	601	16	51	0	0	0	**668**
Total	**39.973**	**1.731**	**2.772**	**33**	**597**	**2**	**45.108**

não há repetição de códigos e todos eles estão ligados apenas a uma IES em que o curso é oferecido. Depois disso, as cinco *tabelas* são unidas[8], de forma que se construa apenas uma *tabela* com dados do CES. Utilizando a nova *tabela* criada a partir da união das *tabelas* do CES de 2010 a 2014, pode-se definir quais vínculos de alunos são considerados como evasão e quais são retenção para três níveis de evasão, a nível de curso, a nível de área de estudo e a nível da IES.

Para a evasão a nível de curso, é considerado evadido o vínculo de aluno que possuir situações de "desvinculado" ou "transferido para outra curso na mesma IES" dentro de cada ano avaliado, pois essas são as únicas situações que indicam uma saída do aluno do curso. As outras situações finais podem indicar saídas do curso, como "formado" ou "falecido", mas não podem ser consideradas evasões, pois o aluno formado concluiu seu curso com sucesso, enquanto que o falecido não pôde continuar os estudos por motivo de óbito. Dessa forma, uma *tabela* de evasão a nível de curso é criada.

Para a evasão ao nível de área de estudo, a definição é diferente, uma vez que, para determinar quais alunos são considerados evadidos de determinada área, deve ser considerado o grupo de cursos que pertencem a área de estudo e não apenas um único curso. Portanto, para o aluno ser considerado evadido, todos os vínculos que o aluno possuir com cursos dessa área devem ser com situações de "transferido" ou "desvinculado". Sendo assim, é verificado se a quantidade de vínculos de um aluno x em determinada área e ano ($Vinc_{x,\text{ano,área}}$) é igual a quantidade de vínculos desse mesmo aluno x com situações de "desvinculado" ($Desv_{x,\text{ano,área}}$) e "transferido" ($Transf_{x,\text{ano,área}}$) ou se é igual a quantidade

[8]O equivalente a um *union all* da linguagem SQL.

de vínculos de "desvinculado" ($Desv_{x,\text{ano,área}}$) que o aluno x possui na mesma área e no mesmo ano, conforme exposto na Equação 3.1, em que, $Evasão_{x,\text{ano,área}}$ recebe 1 se o aluno x for considerado evadido da área de estudo no ano ou 0 caso contrário.

$$Evasão_{x,\text{ano,área}} = \begin{cases} 1, \text{se} & Vinc_{x,\text{ano,área}} = Desv_{x,\text{ano,área}} + Transf_{x,\text{ano,área}} \\ 1, \text{se} & Vinc_{x,\text{ano,área}} = Desv_{x,\text{ano,área}} \\ 0, & \text{caso contrário} \end{cases} \tag{3.1}$$

A necessidade de se comparar duas quantidades de vínculos, $Desv_{x,\text{ano,área}} + Transf_{x,\text{ano,área}}$ e apenas $Desv_{x,\text{ano,área}}$, vem da característica de preenchimento do CES, pois não é permitido que um aluno possua apenas um vínculo de "transferido" em determinado ano. Como essa situação descreve uma transferência dentro da IES, é necessário que o aluno possua pelo menos mais um vínculo em um curso distinto [59], o curso de destino da transferência.

A evasão a nível da IES é muito similar a evasão por área, uma vez que, se faz necessário verificar se o aluno possui apenas vínculos de evasão ("desvinculado" e "transferido") em todos os cursos da instituição. A diferença ocorre no agrupamento, isto é, verifica-se a quantidade de vínculos do aluno na instituição ao invés de fazê-lo na área de estudo. Isso pode ser visto na Equação 3.2, em que, $Evasão_{x,\text{ano,IES}}$ recebe 1 se o aluno for considerado evadido ou 0 caso contrário; $Vinc_{x,\text{ano,IES}}$ é a quantidade de vínculos do aluno x por ano em determinada IES; $Desv_{x,\text{ano,IES}}$ é a quantidade de vínculos do aluno x por ano na IES que são iguais a desvinculado; $Transf_{x,\text{ano,IES}}$ é a quantidade de vínculos de "transferido" que o aluno x possui em determinado ano na IES.

$$\text{Evasão}_{x,\text{ano,IES}} = \begin{cases} 1, \text{se} & Vinc_{x,\text{ano,IES}} = Desv_{x,\text{ano,IES}} + Transf_{x,\text{ano,IES}} \\ 1, \text{se} & Vinc_{x,\text{ano,IES}} = Desv_{x,\text{ano,IES}} \\ 0, & \text{caso contrário} \end{cases} \tag{3.2}$$

Verificando as definições para se rotular um vínculo como evasão ou não, percebe-se que as características dos cursos ABI podem fazer com que vínculos que não são considerados evasão, sejam rotulados erroneamente. A situação em que o aluno se transfere de um curso ABI para um curso de formação, é um desses casos. Enquanto que, em cursos sem Área Básica de Ingresso um aluno transferido pode ser considerado uma evasão daquele curso, este não é o caso para um curso ABI, em que o aluno está seguindo seu fluxo normal. Dessa forma, retira-se da *tabela* os vínculos que possuem a situação de "transferido para outro curso na mesma IES" e são ligados a cursos ABI.

Tabela 3.9: Número de vínculos de alunos ingressantes por situação e área geral de estudo nos cursos de graduação presencial, exceto os ABI, da UnB de 2010 a 2014.

	Situação de vínculo						
Área geral	Cursando	Trancado	Desvinculado	Transferido	Formado	Falecido	**Total**
1	8.339	492	689	22	184	0	**9.726**
2	2.439	130	206	5	62	0	**2.842**
3	10.442	453	638	1	176	2	**11.712**
4	3.331	130	219	0	168	0	**3.848**
5	5.109	152	305	1	3	0	**5.570**
6	1.531	50	128	0	2	0	**1.711**
7	4.863	198	322	1	2	0	**5.386**
8	601	16	51	0	0	0	**668**
Total	**36.655**	**1.621**	**2.558**	**30**	**597**	**2**	**41.463**

Um outro caso em que as características dos cursos ABI pode ocasionar a rotulação errônea de vínculos é em relação a evasão no nível de área de estudo. Apenas um dos cursos ABI da UnB, dentre dezessete, possui área de estudo nos dados do CES ao longo dos 5 anos analisados. Portanto, não é possível identificar de qual área de estudo um aluno vinculado a um curso ABI evadiu. Por essa razão, todos os vínculos associados aos dezessete cursos ABI serão desconsiderados da análise para o nível de evasão de área de estudo. A Tabela 3.9 mostra as mesmas informações da Tabela 3.8 com a retirada dos vínculos de alunos ligados a cursos ABI. Pode-se ver que os cursos ABI da UnB possuíam três diferentes classificações de grande área de estudo. Ou estavam com a área em branco, ou classificados como "Programas ou cursos gerais" ou "Engenharia, Produção e Construção".

Com a definição de evasão para os três níveis e a criação das *tabelas* com dados do CES para cada nível, o próximo passo é agregar os dados do ENEM às *tabelas*. Conforme explicado anteriormente, os dados do ENEM são distribuídos em cinco *tabelas*, uma para cada ano (2010 a 2014). As *tabelas* dos cinco anos serão unidas criando uma única *tabela* do ENEM com as informações de todos os anos. A junção das *tabelas* agregadas do CES e ENEM é feita de forma que se utilize a última informação disponível do aluno no ENEM sem que ela seja posterior à informação do aluno no CES. Ou seja, para um aluno que aparece no ano de 2013 no CES e fez o exame do ENEM em 2010, 2012 e 2014, as informações do ENEM devem ser referentes a 2012, o último ano que possui informações desse aluno não sendo posterior à informação do CES.

Apesar do ENEM servir de subsídio para o ingresso em algumas Instituições de Ensino Superior, para este estudo, o aluno não necessariamente precisa ter ingressado na IES através do ENEM, ou seja, o aluno pode ter ingressado por vestibular, mas realizado o

Tabela 3.10: Número de vínculos de alunos ingressantes por situação e área geral de estudo que não participam do estudo..

	Situação de vínculo						
Área geral	Cursando	Trancado	Desvinculado	Transferido	Formado	Falecido	**Total**
1	4.427	281	316	18	175	0	**5.217**
2	1.430	93	102	1	60	0	**1.686**
3	4.860	228	295	0	174	2	**5.559**
4	1.551	48	86	0	163	0	**1.848**
5	2.056	60	129	1	3	0	**2.249**
6	633	12	47	0	2	0	**694**
7	1.679	58	94	0	2	0	**1.833**
8	248	6	18	0	0	0	**272**
Total	**16.884**	**786**	**1.087**	**20**	**579**	**2**	**19.358**

exame em algum tempo anterior ao seu ingresso para ser considerado na análise. Conforme explicado anteriormente, apenas observações de alunos que realizaram todas as provas de pelo menos uma edição do ENEM no período entre 2010 e 2014 serão utilizados. Dessa forma, alunos da UnB que não realizaram as provas (não basta ter se inscrito no exame), ficam fora da análise.

Existe, portanto, um grupo de alunos da UnB que não são analisados neste trabalho. A Tabela 3.10 mostra como esses alunos se distribuem de acordo com as situações de vínculo e áreas de estudo. Nota-se que sua distribuição é similar a vista na Tabela 3.9. Os dados provenientes do exame servem como informação, anterior ao ingresso no ensino superior, do desempenho acadêmico e de características socioeconômicas dos alunos selecionados no CES. Essas informações serão utilizadas como atributos preditivos para os modelos de classificação para ajudar a determinar se um estudante evade ou não. Como qualquer pessoa pode fazer a prova do ENEM a qualquer tempo, não associar a forma de ingresso do aluno ao exame permite analisar uma quantidade maior de alunos da IES.

Como os questionários feitos aos inscritos do ENEM foram alterados ao longo do tempo, foi necessário realizar um tratamento dos dados para que os atributos pudessem ser úteis para os classificadores em qualquer ano. Para criar a *tabela* unida dos dados do ENEM, não é necessário filtrar informações nas *tabelas*, pois elas possuem as informações de todos os inscritos no exame. No entanto, é necessário padronizar os atributos de anos diferentes para unir as *tabelas* de anos distintos. Toma-se como referência as *tabelas* e questionários de 2013 e 2014[9], pois elas refletem a evolução do questionário ao longo dos anos e por elas serem praticamente iguais em termos dos atributos que possuem. Dessa forma, a preparação das *tabelas* de 2013 e 2014 consiste apenas em selecionar os

[9] `https://github.com/lucke71/Classify_dropout/tree/master/Dicionários%20ENEM`

atributos. São selecionados 133 em cada ano, dos quais 76 são perguntas do questionário. Os outros 57 atributos incluem 10 notas (4 notas das áreas de conhecimento, 5 notas de competências na redação e a nota da própria redação); 3 de identificação (CPF, ano da informação e identificação de inscrição); 26 indicadores de necessidades especiais para realizar o exame ou de alguma deficiência do inscrito; 2 atributos relacionados a realização do exame (município onde o inscrito fez a prova e se o inscrito requereu certificação de ensino médio) e 10 atributos com características dos inscritos:

- indicador de inscrição em unidade hospitalar[10];
- município de residência;
- idade;
- sexo;
- nacionalidade;
- município de nascimento;
- indicador de conclusão do ensino médio;
- ano que concluiu o ensino médio;
- estado civil e
- cor/raça.

E ainda 6 atributos relacionados à escola do inscrito, no caso de ele estar cursando o ensino médio:

- município da escola;
- dependência da escola (federal, estadual, municipal, privada) ;
- localização da escola (urbana, rural);
- situação de funcionamento da escola (em atividade, paralisada, extinta, extinta em anos anteriores);
- tipo da escola (pública, privada) e
- tipo de ensino (regular, de jovens e adultos, especial).

[10]Inscritos que estão internados e necessitam realizar o exame em unidade hospitalar.

Para a *tabela* de 2012, tentou-se trazer os mesmos atributos, no entanto alguns não existem na edição de 2012 do exame. Nacionalidade, município de nascimento e as questões de números 40 a 53 do questionário sócio econômico de 2014 não fizeram parte das informações coletadas dos inscritos para o exame em 2012, como pode ser visto no dicionário de 2012 do ENEM. Para resolver a falta da informação da nacionalidade dos alunos, empregou-se o mesmo atributo de nacionalidade, mas advindo da *tabela* do CES. Já para o município de nascimento, não existe tal informação nas *tabelas* do CES e portanto, o atributo fica em branco para 2012. As perguntas do questionário sócio-econômico são relacionadas à possível atividade remunerada dos inscritos e aos cursos que os inscritos frequentam ou já frequentaram. Há ainda uma alteração na questão de número 22 (q22) do questionário de 2012 (questão de mesmo número no questionário de 2014)[11]. A resposta para essa questão permite duas respostas "A - Sim" ou "B - Não", enquanto que em 2014, as possibilidades de respostas são três:

- A - Sim, estou trabalhando;
- B - Sim, já trabalhei, mas não estou trabalhando e
- C - Não, nunca trabalhei.

A pergunta para a questão é a mesma em ambos os anos: "Você exerce ou já exerceu atividade remunerada?". Portanto, a substituição da letra "B" por "C" no questionário de 2012, serve como uma adaptação para a questão de 2014. Nas *tabelas* de 2011 e 2010 existem mais diferenças em relação às *tabelas* de 2013 e 2014 e serão detalhadas a seguir.

Na *tabela* de 2011, assim como em 2012, não existem os atributos de nacionalidade e município de nascimento, dessa forma, as mesmas soluções empregadas na *tabela* de 2012 são empregadas para 2011. Além disso, 11 perguntas do questionário sócio-econômico de 2014 não são feitas aos inscritos. São elas: q003; q004; q007 a q021; q024 a q027; q029; q036 a q040; q062; q063; q067; q069 e q075. Ainda se faz necessário ajustar 18 outros atributos do questionário. Nove deles, q01; q09; q10; q11; q12; q13; q23; q24 e q27 (questionário 2011 do ENEM) são transformados para atributos numéricos. Essa transformação é necessária para que os atributos de 2011 sejam da mesma categoria dos de 2014. Entre os nove atributos, seis deles (q09 a q13 e q23) também requerem mudanças de valores, pois os valores em branco se apresentam de formas diferentes, em alguns casos como espaços em outros como pontos (.). Dessa forma, além de mudar a categoria desses seis atributos para numéricos, também foi alterada a forma como eles apresentavam os valores faltantes.

[11] As questões de 2013 e 2014 possuem o formato de tês dígitos (qxxx), enquanto que nos outros anos o formato é de dois dígitos (qxx).

As últimas alterações necessárias para a *tabela* de 2011 são feitas nos atributos q44; q45; q46; q47 a q51 (questionário 2011 do ENEM). Essas perguntas são equivalentes as q065; q066; q068; q070 a q074 do questionário de 2014 e dizem respeito ao abandono do ensino médio regular. Em ambos os anos as questões indagam sobre a relevância de fatores contribuintes diferentes para o abandono do ensino médio regular. No entanto em 2011, as opções de respostas são apresentadas em níveis, de 0 a 5, em que 0 indica a pouca relevância do fator no abandono. Em 2014, as respostas possíveis são dicotômicas, ou seja, 'Não' para o caso de o fator não ter contribuído ou 'Sim', caso tenha contribuído. A adaptação realizada foi a de dividir as seis respostas possíveis do questionário de 2011 em duas opções, portanto quem respondeu que o fator contribuinte teve relevância de 0 a 2 teve a resposta alterada para 'Não', enquanto quem respondeu de 3 a 5 teve a resposta alterada para 'Sim'.

Para a *tabela* de 2010 os ajustes são similares aos da *tabela* de 2011. Mais uma vez, não foram coletadas as informações de nacionalidade e município de nascimento e a mesma solução empregada anteriormente foi aplicada. Existem 32 questões do questionário de 2014 que não existem em seu correspondente de 2010. Elas são: q003; q004; q007 a q021; q024 a q027; q029; q036 a q040; q062; q063; q067; q069 e q075. As perguntas q02 e q03 do questionário de 2010 possuem praticamente as mesmas respostas de suas correspondentes no questionário de 2014 (q001 e q002) exceto que em 2014 as escolaridades de 'mestrado' e 'doutorado' foram substituídas por 'pós-graduação'. Além de agregar as repostas de 'mestrado' e 'doutorado' as respostas necessitaram ser trocadas de ordem para serem compatíveis com as de 2014.

A pergunta q06 de 2010, cuja correspondente é a q005 de 2014 possui mais opções de resposta em 2014, dessa forma a transformação realizada foi apenas de rearranjar as letras das respostas de 2010 para serem compatíveis ao questionário de 2014. A próxima questão que necessitou de ajuste foi a q08 de 2010, cuja questão equivalente em 2014 é a q022. Mais uma vez, existem respostas a mais em 2014, portanto só se faz necessário um rearranjo da ordem das opções de resposta em 2010. A q30 e q33, cujas equivalentes em 2014 são q032 e q035, respectivamente, possuem seis opções de resposta em comum e na mesma ordem. O único ajuste necessário é transformar duas respostas de 2010 em informações em branco, pois elas não existem em 2014. O último ajuste a ser feito na *tabela* de 2010 é similar ao realizado nas questões finais de 2011. As questões q45 a q52 apresentam respostas em níveis de 0 a 5, enquanto que as questões equivalentes em 2014 (q065; q066; q068 e q070 a q074) apresentam respostas dicotômicas. Sendo assim, é feito o mesmo tratamento realizado na *tabela* de 2011 para a *tabela* de 2010.

Como apontado na seção anterior, a *tabela* referente ao ano de 2010 do ENEM possui informações faltantes. São 14.478 observações que não possuem valores em nenhuma das

questões do questionário socioeconômico e nos atributos relacionados ao inscrito, como cor/raça, estado civil e deficiências. A maioria delas também não possuem CPF ou notas nas provas. As observações apresentam um comportamento estranho de informações faltantes, isto é, nem todos os atributos estão faltantes, mas não há informações do questionário socioeconômico. Em análise com técnicos do Inep sobre o caso, foi sugerido retirar as observações da *tabela* final, pois é possível que elas sejam o produto de testes que não foram resolvidos à época e acabaram entrando na base final.

Uma vez que todas as adaptações necessárias aos questionários são realizadas, pode-se unir as *tabelas* de 2010 a 2014, formando uma única *tabela* com informações de todos os anos do ENEM e realizar a junção com a *tabela* agregada do CES. A junção é realizada através do CPF dos alunos, o único atributo em comum em ambas as *tabelas*. Na junção das *tabelas*, é aplicada a regra de utilizar a última informação disponível dos alunos no ENEM sem que ela seja posterior à informação do CES. Além disso, não foram utilizadas as informações dos alunos que não são encontradas em nenhum dos anos do ENEM, e na união das *tabelas* do ENEM foram retiradas as observações de inscritos sem notas em cada ano, para que ao juntar com a *tabela* do CES não existam observações faltantes.

A retirada de observações no momento da formação da *tabela* agregada do ENEM é importante, pois se ocorresse após a junção, resultariam em menos observações utilizadas. Por exemplo, um aluno do CES de 2014 possui informações nos anos de 2012 e 2011 do ENEM, sendo que em 2011 existe a informação das notas e em 2012 não. Essa linha seria descartada se as observações faltantes não fossem retiradas no momento da formação da *tabela* agregada do ENEM, pois na junção das *tabelas* somente a última informação (ano de 2012) seria utilizada. A retirada dos alunos que se inscreveram no ENEM e não realizaram provas se justifica, pois não é possível aferir o desempenho acadêmico desses alunos. Alternativamente, valores poderiam ser estimados para os alunos, no entanto esse não é o foco deste trabalho. Ainda, estudos que visam estimar o desempenho de alunos geralmente utilizam alguma medida de desempenho anterior do mesmo estudante como atributo preditivo [66, 67]. Dessa forma, obtém-se três *tabelas* finais (uma para cada nível de evasão) para a etapa de modelagem do CRISP-DM.

A Tabela 3.11 mostra a quantidade de vínculos que existem em cada uma das *tabelas* criadas para os níveis de evasão. Percebe-se que a quantidade de observações total diminuiu em todas as *tabelas* quando comparado ao total de observações das Tabelas 3.8 e 3.9. Essa diminuição se deve à junção das bases do CES e ENEM. A quantidade de vínculos diminuiu por mais da metade, mostrando que muitos dos alunos que ingressaram na UnB entre 2010 e 2014 não se inscreveram no ENEM ou faltaram a prova ao longo do mesmo período de tempo. Nota-se também que a *tabela* para evasão a nível da área do curso apresenta menos observações no total. Isso ocorre, pois todos os vínculos de alunos

ligados a cursos ABI foram descartados nessa *tabela*.

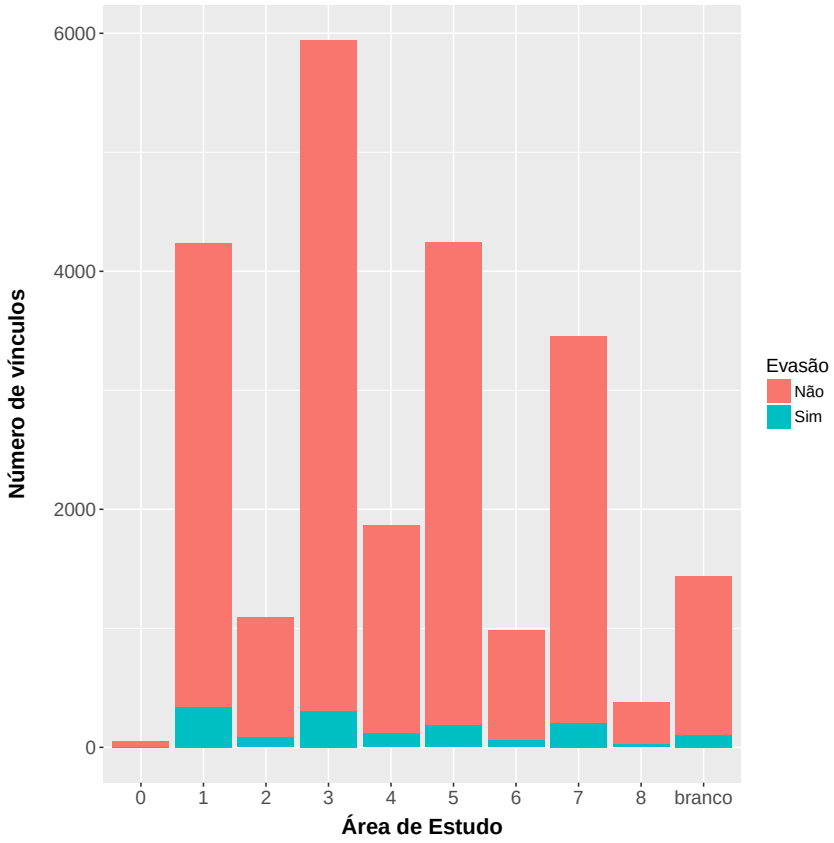

Figura 3.3: Quantidade de vínculos por evasão e área de estudo.

Outra observação importante é que, o número de alunos rotulados como evasão é diferente nas *tabelas* de evasão por nível de curso e nível de IES, apesar de ambas possuírem a mesma quantidade de observações. Isso se deve às diferentes características desses níveis de evasão. Na evasão a nível de IES, a evasão só se dá quando todos os vínculos do aluno na IES são de 'desvinculado' ou 'transferido', enquanto que no nível de curso, basta que o aluno possua um vínculo com uma dessas situações no curso. Dado que existem alunos com mais de um vínculo na UnB é lógico que, quanto maior o nível de agregação, menor a quantidade de vínculos rotulados como evasão. Além disso, uma das características dos dados do CES e, por consequência da rotulação dos vínculos em evasão e retenção é que um mesmo aluno pode ser considerado ao mesmo tempo evasão e retenção nas *tabelas* finais. Na *tabela* de evasão a nível de área de estudo são 332 alunos distintos com dupla rotulação, na *tabela* a nível de curso, são 455 alunos distintos e na *tabela* a nível de IES são 193 alunos com essa característica. Nesses casos, os atributos relacionados ao curso a que o aluno está vinculado auxiliarão os modelos a classificarem a observação.

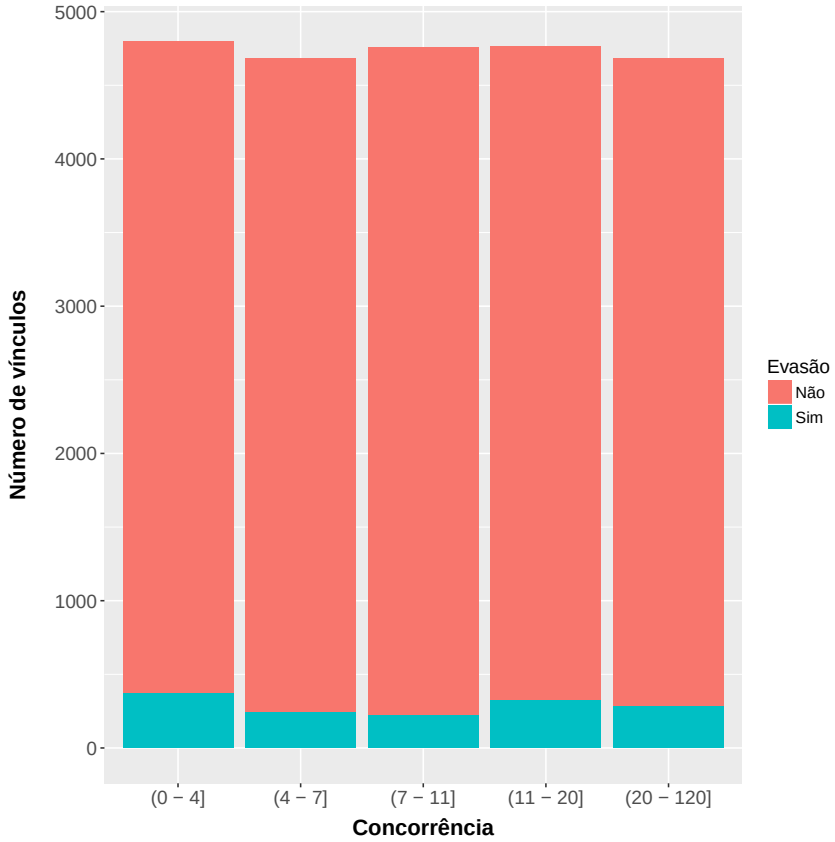

Figura 3.4: Quantidade de vínculos por evasão e concorrência.

O próximo passo é realizar uma conexão entre o *software* R e o servidor SQL, conforme explicado anteriormente. Dessa forma, pode-se extrair as *tabelas* criadas no servidor SQL para o R sem exportá-las para um outro formato antes. Com os dados importados, alguns procedimentos são realizados, como indicar quais atributos não são quantitativos e padronizar seus valores para que o *software* não diferencie, por exemplo, 'A' de 'A_' ou ' A ' (com espaços). Além da limpeza dos valores, também é necessário tratar os atributos que possuem algum valor em branco. Os dados do CES e ENEM não possuem informações faltantes, mas algumas peguntas no questionário socioeconômico forçam atributos a ficarem em branco. Por exemplo, a questão 22 de 2014 (q022) indaga o inscrito se ele já exerceu

Tabela 3.11: Número de vínculos e percentual de alunos nas *tabelas* agregadas de CES e ENEM por nível de evasão e rótulo dos vínculos.

———	Nível de evasão					
Rótulo	Curso	%	Área de estudo	%	IES	%
Evasão	1.463	6,18	1.256	5,94	1.211	5,12
Retenção	22.229	93,82	19.911	94,06	22.481	94,88
Total	**23.692**	**100,00**	**21.167**	**100,00**	**23.692**	**100,00**

atividade remunerada, enquanto que a questão 40 (q040) pergunta ao inscrito com que idade ele começou a exercer atividade remunerada. Fica claro que se a resposta para a q022 for negativa, o valor da q040 ficará em branco. Nesses casos, optou-se por preencher um valor comum para que os classificadores não deixassem de utilizar a observação. Nesses casos, a *string* "branco" foi preenchida.

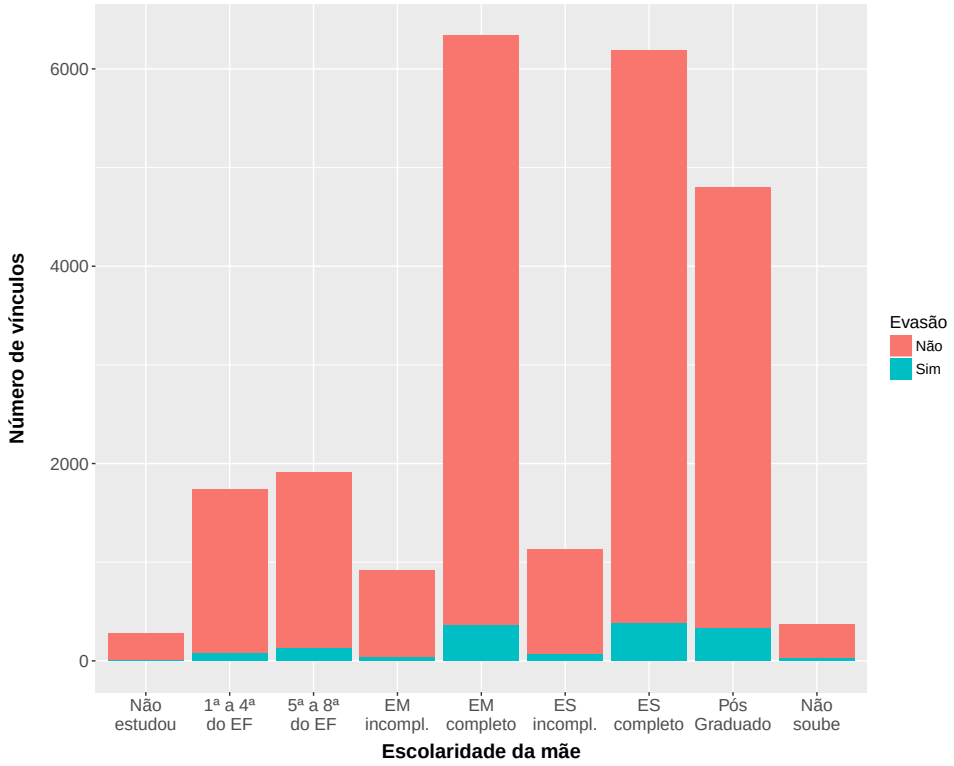

Figura 3.5: Quantidade de vínculos por evasão e escolaridade da mãe.

Nas *tabelas* de evasão por níveis existem quatro atributos de código de município, todos relacionados ao aluno. Município de residência, de sua escola, de nascimento e de realização da prova do ENEM. Existem mais de cinco mil municípios no Brasil e quatro atributos distintos que possam conter essa quantidade de informação pode fazer com que os classificadores se confundam. A informação de onde o aluno nasceu ou fez a prova pode ser considerada não relevante para ajudar a classificar um aluno que evadiu no ensino superior, mas a informação de que ele nasceu em um município distinto daquele em que ele reside quando fez a prova ou até que esse aluno realizou o exame do ENEM em um município que não o que ele reside pode ser uma informação útil para o classificador. Dessa forma foram criados atributos para todas as seis combinações possíveis de diferença entre os municípios. Além disso, para não descartar a informação sobre a localização do aluno, foram criados quatro atributos indicando a Unidade da Federação a que o município pertence.

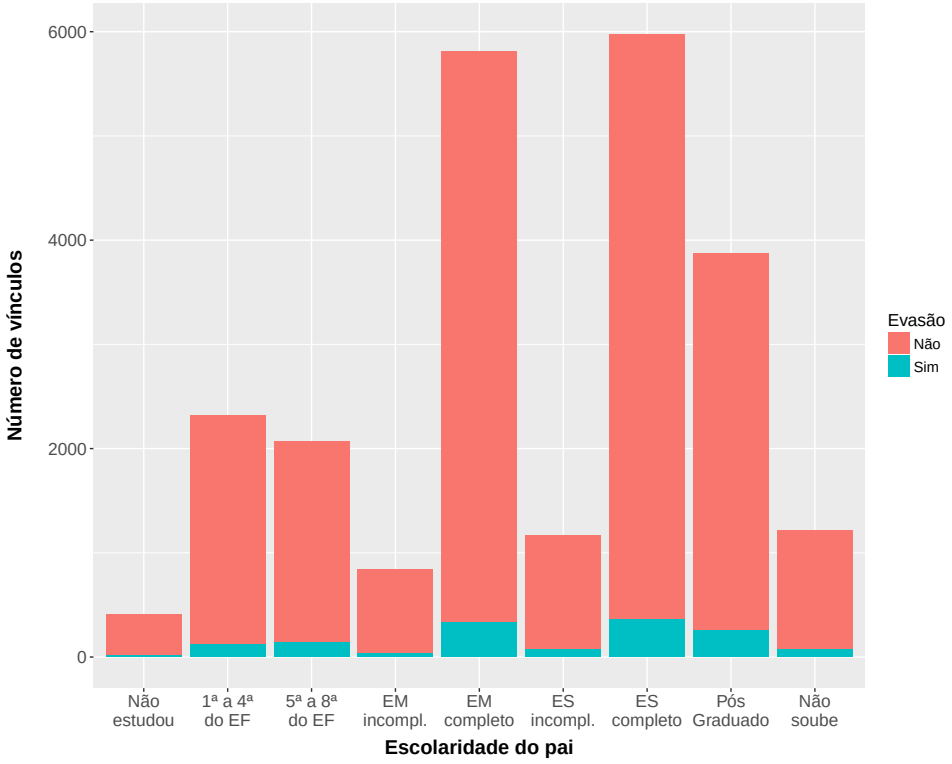

Figura 3.6: Quantidade de vínculos por evasão e escolaridade do pai.

Outro atributo que necessita de tratamento é o ano de conclusão do ensino médio, pois fica em branco se o aluno não tiver concluído. Neste caso, foi criado um atributo que indica se o aluno concluiu o ensino médio, ou não, baseado no atributo 'ano de conclusão' ter, ou não, uma resposta. No atributo numérico, 'ano de conclusão', foi inserido o valor '0' para que os classificadores possam utilizar as observações.

O último procedimento antes de começar o treinamento dos classificadores foi remover os atributos que possuem pouca variabilidade e dessa forma podem não contribuir para realizar as classificações. Foi considerada pouca variabilidade quando a razão entre a frequência do valor mais comum de um atributo e o segundo mais comum apresentar valor inferior a 97/3. Assim sendo, as *tabelas* estão finalizadas e prontas para que os classificadores sejam treinados.

Antes de seguir para a etapa de modelagem dos dados, é realizada uma breve análise descritiva dos dados. Foram escolhidos alguns atributos para mostrar suas relações com a classe de interesse, a evasão. A escolha dos atributos foi baseada no conhecimento de técnicos do Inep, familiarizados com o negócio, que afirmam que esses atributos são importantes para descrever alunos evadidos. Alguns atributos como o desempenho dos alunos em avaliações anteriores ao ingresso também são vistas em alguns estudos como [2, 29, 55], outros como o semestre de ingresso e a situação de conclusão foram incluídos

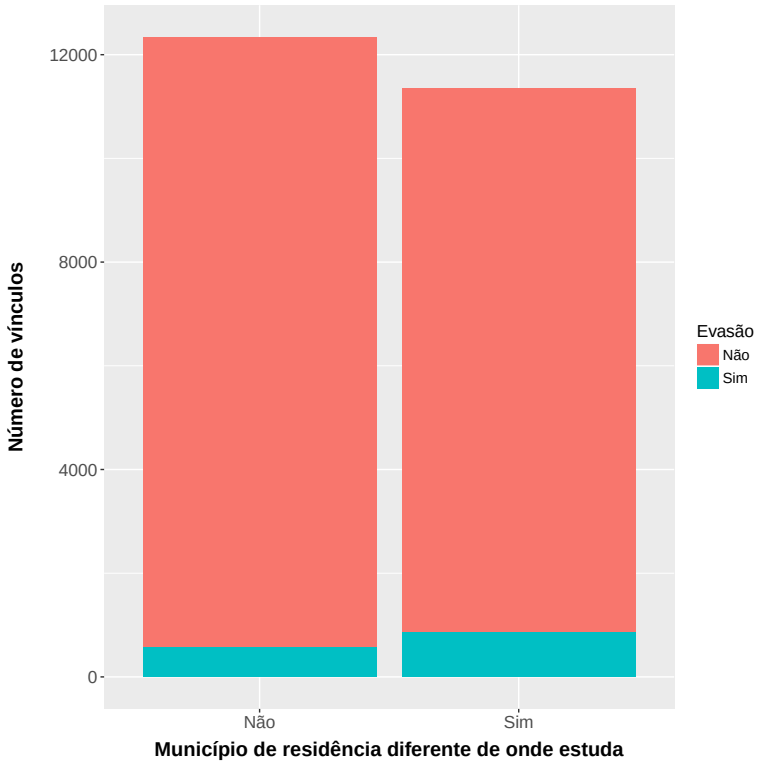

Figura 3.7: Quantidade de vínculos por evasão e indicador de morar em município distinto da escola onde estudou no Ensino Médio.

baseados nos resultados de modelagens preliminares.

Como as *tabelas* criadas para cada nível possuem características muito próximas, afinal a única diferença entre elas é na rotulação das observações quanto a evasão, as análises foram realizadas apenas na *tabela* de evasão a nível de curso. Apesar de não haver uma clara diferença na análise descritiva dos atributos entre as *tabelas* de níveis de evasão, não é possível identificar se não há diferença de desempenho entre os modelos criados para *tabelas* distintas.

Começando a análise pela área de estudo, a Figura 3.3 mostra quantidade de vínculos evadidos por grande área de estudo. A área que mostra a maior proporção de evadidos é a área 1 (Educação), sendo que a área 3 (Ciências Sociais, Negócios e Direito) é a área que possui a maior quantidade de vínculos. Aparentemente, não há uma concentração de evasão em nenhuma das áreas.

A concorrência, definida aqui pela razão entre a quantidade de candidatos no curso e a quantidade de vagas oferecidas, foi dividida em categorias. As categorias foram criadas a partir de 5 quantis (0 - 0,2; 0,2 - 0,4; 0,4 - 0,6; 0,6 - 0,8; 0,8 - 1,0). A Figura 3.4 mostra que o intervalo com maior proporção de evadidos é o de 0 a 4 candidatos por vaga. Espera-se que quanto maior a concorrência, menor seja a taxa de evasão, no entanto, essa

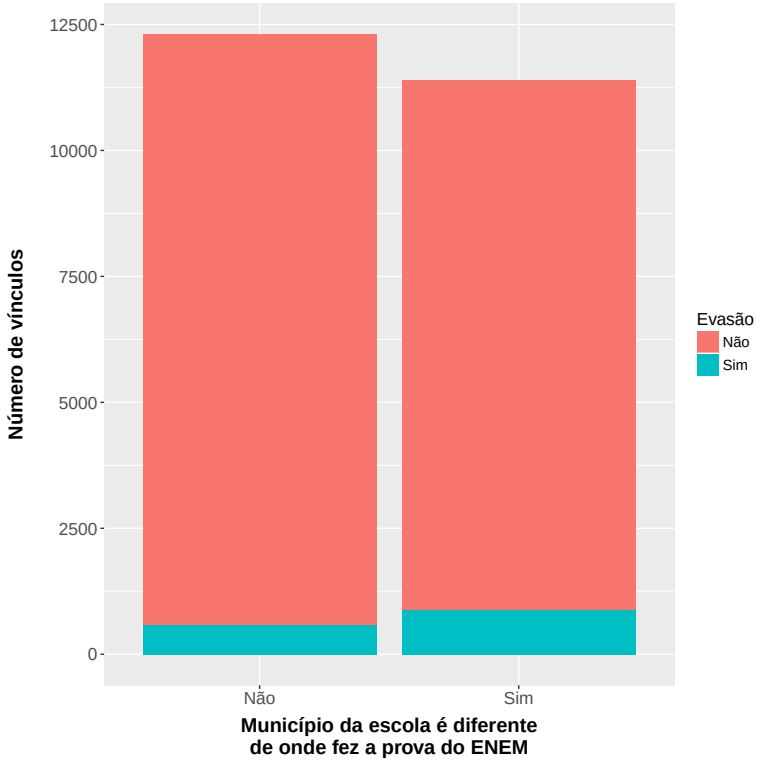

Figura 3.8: Quantidade de vínculos por evasão e indicador de que escola e local de prova do ENEM ficam em municípios distintos.

expectativa não é atingida observando a Figura 3.4, pois a proporção de evadidos aumenta nos dois quantis mais elevados (11 a 20 e 20 a 120).

A Figura 3.5 mostra a quantidade de vínculos por evasão e escolaridade da mãe no momento da realização da inscrição do ENEM. A maior proporção de alunos evadidos ocorre quando a escolaridade da mãe é de pós-graduada, seguida de ensino superior completo. Essas informações são contra intuitivas, uma vez que espera-se que a evasão esteja atrelada a aspectos negativos do indivíduo, no caso da escolaridade de seus pais, no entanto, não é esse o caso. O mesmo comportamento pode ser observado na Figura 3.6, em que quase não há diferença na distribuição dos alunos evadidos.

Como descrito anteriormente, os atributos de município de residência e de município de escola do Ensino Médio foram utilizados para se criar um atributo que indica se há diferença entre eles para os alunos. A Figura 3.7 mostra as quantidades e proporções de alunos evadidos a depender do valor do atributo criado. A quantidade de alunos que possuem valor "Sim" para o atributo é um pouco menor que os que possuem "Não". Em relação a proporção, os alunos evadidos aparecem em maior proporção quando os alunos residem em municípios distintos dos que estudaram no Ensino Médio.

De forma similar, os atributos de município da escola e de local de prova do ENEM

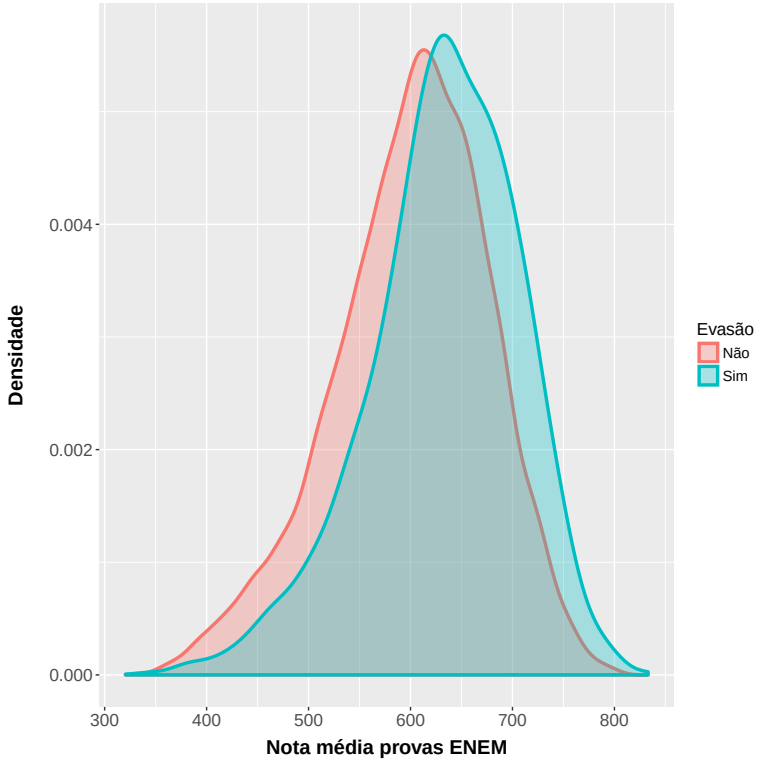

Figura 3.9: Distribuição das notas médias dos alunos nas provas de acordo com o rótulo de evasão.

foram utilizados para formar um atributo que indica quando eles são diferentes para cada aluno. A Figura 3.8 mostra as proporções de evasão para cada valor do atributo. A figura mostra um resultado muito similar ao da Figura 3.7. Os alunos que fizeram as provas do ENEM em municípios distintos dos que eles vão à escola, possuem uma proporção maior de evasão.

As Figuras 3.9 e 3.10 mostram as distribuições das notas dos alunos. A Figura 3.9 mostra a média dos alunos nas quatro provas, enquanto que a Figura 3.10 mostra a nota na redação. A interpretação dos gráficos é a mesma. Quanto maior a nota do aluno, maior a proporção de alunos rotulados como evasão. Isso pode indicar que alunos com notas altas possuem maior possibilidade de conseguir ingressar nos cursos que desejam, evadindo de um curso que podem ter ingressado como segunda opção.

A Figura 3.11 mostra a quantidade de vínculos com IES distintas que cada aluno possui em todo o Ensino Superior e a proporção de evasão em cada valor do atributo. A maior parte dos alunos possui vínculo com apenas uma IES, no entanto, os que possuem vínculo com 2 IES aparentam possuir uma proporção maior de alunos evadidos.

A Figura 3.12 traz a informação do semestre de ingresso dos alunos na UnB e a proporção de evadidos em cada valor do atributo. Pode-se notar que a proporção de

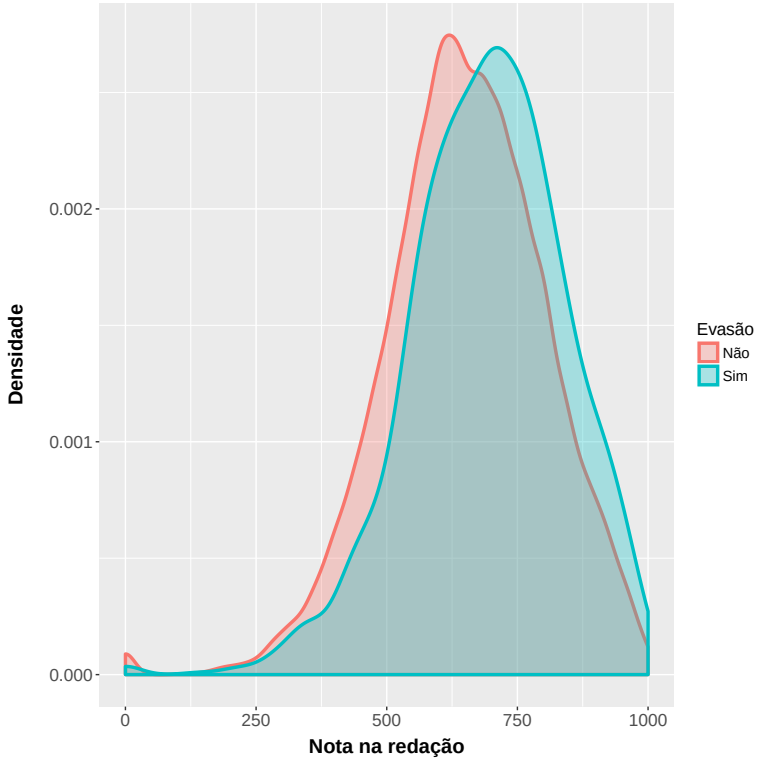

Figura 3.10: Distribuição das notas dos alunos em redação de acordo com o rótulo de evasão.

evadidos no primeiro semestre é maior que a do segundo semestre. Percebe-se também que a quantidade de ingressos no primeiro semestre é superior aos ingressos no segundo semestre.

A Figura 3.13 mostra a situação de conclusão no Ensino Médio que os alunos possuíam ao realizar a inscrição do ENEM. Alunos que já haviam concluído o Ensino Médio ao ingressar na UnB, possuem uma proporção de evasão maior que o dos outros valores do atributo. A figura mostra que alunos que ainda vão concluir o Ensino Médio, isto é, alunos que estão seguindo o fluxo, quando fazem o ENEM possuem uma taxa de evasão menor.

A Figura 3.14 mostra o tipo de escola de Ensino Médio que os alunos estavam estudando no momento da inscrição no ENEM. Nota-se que a quantidade de alunos provenientes de escolas públicas é menor que o de escolas particulares e que a proporção desses alunos que evade também é menor. Existe uma quantidade bastante grande de alunos que não responderam à pergunta. Isso se deve ao fato de eles não estarem mais cursando o Ensino Médio, já que a pergunta se limita a esse grupo de alunos. Mais uma vez observa-se que o grupo de alunos que já havia concluído o Ensino Médio possui uma proporção de evasão maior que a do grupo de alunos que ainda não concluiu.

A Figura 3.15 mostra a situação dos alunos quanto a suas ocupações. Mais uma vez o

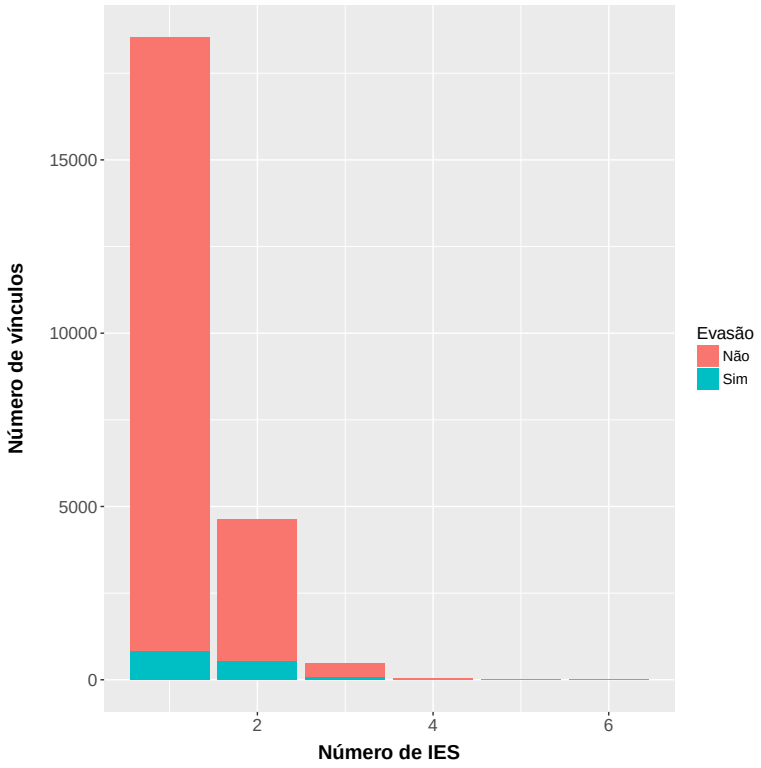

Figura 3.11: Quantidade de vínculos de alunos por evasão e número de IES vinculadas.

resultado é contra intuitivo, pois espera-se que alunos que exercem algum tipo de atividade empregatícia evadam mais que alunos que apenas estudam. Analisando a Figura 3.15 não é isso que se observa. Sendo que a maior proporção de evasão ocorre com os alunos que nunca trabalharam.

Nesta etapa foi detalhada a escolha das observações e a criação dos atributos que fazem parte do estudo. Em seguida, foi explicada a metodologia de rotulação das observações em evasão e não evasão nas *tabelas* do CES. Também foi detalhada a forma com que a junção entre as bases do CES e ENEM ocorre. Vale destacar que, a partir da preparação dos dados do ENEM para a junção foram encontradas observações sem preenchimento, que não deveriam existir. Também foi realizada uma breve análise descritiva dos dados mostrando a relação de alguns atributos com a classe de interesse. Por último foram descritos os procedimentos realizados com os dados no ambiente R, preparando as *tabelas* com os diferentes níveis de enfoque dos dados preparados para que se possa seguir com a etapa de modelagem do CRISP-DM.

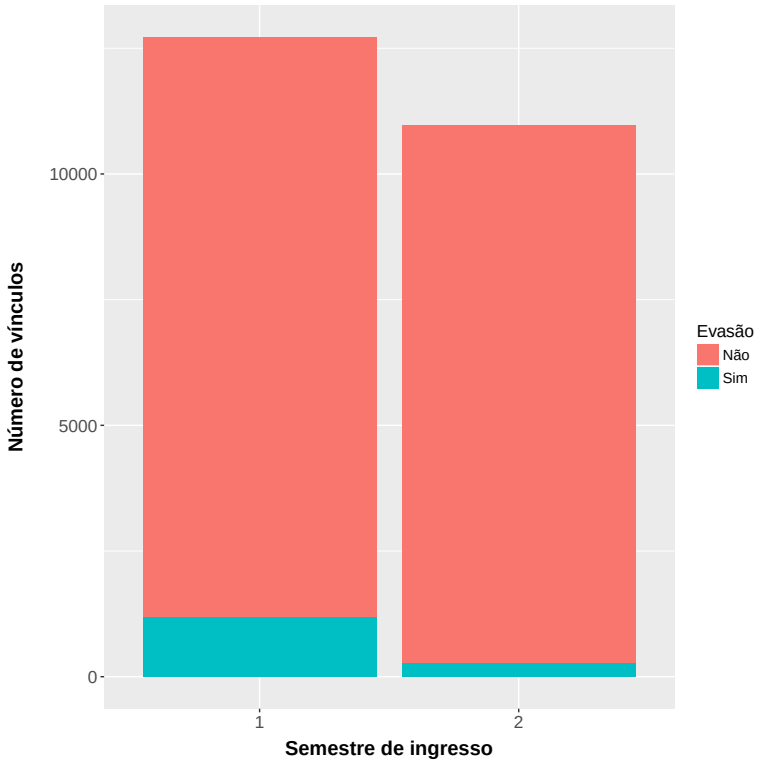

Figura 3.12: Quantidade de vínculos de alunos por evasão e semestre de ingresso.

3.4 Modelagem

Nesta etapa, várias técnicas de modelagem foram empregadas. Como muitos modelos são dependentes do formato dos dados, fez-se necessário voltar ao estágio de preparação dos dados inúmeras vezes. A modelagem é dividida em:

- seleção da técnica de modelagem;
- geração do *design* do teste e
- criação dos modelos.

Para realizar a modelagem, utilizou-se o *software* R, que possui grande quantidade de pacotes disponíveis. O pacote *caret*[12] foi empregado, pois padroniza a sintaxe de treinamento e predição de modelos de classificação e possui todos os algoritmos que se pretende utilizar: CART, C5.0, regressão logística, redes neurais artificiais e *Naive Bayes*. Além disso, ele possui funcionalidades para treinar os algoritmos com diferentes configurações de parâmetros de forma automática ou manual (*tuning*); implementações para realizar separação de dados em treinamento e teste; ferramentas de pré-processamento, seleção de atributos e estimação da importância dos atributos preditivos.

[12] `http://topepo.github.io/caret/index.html`

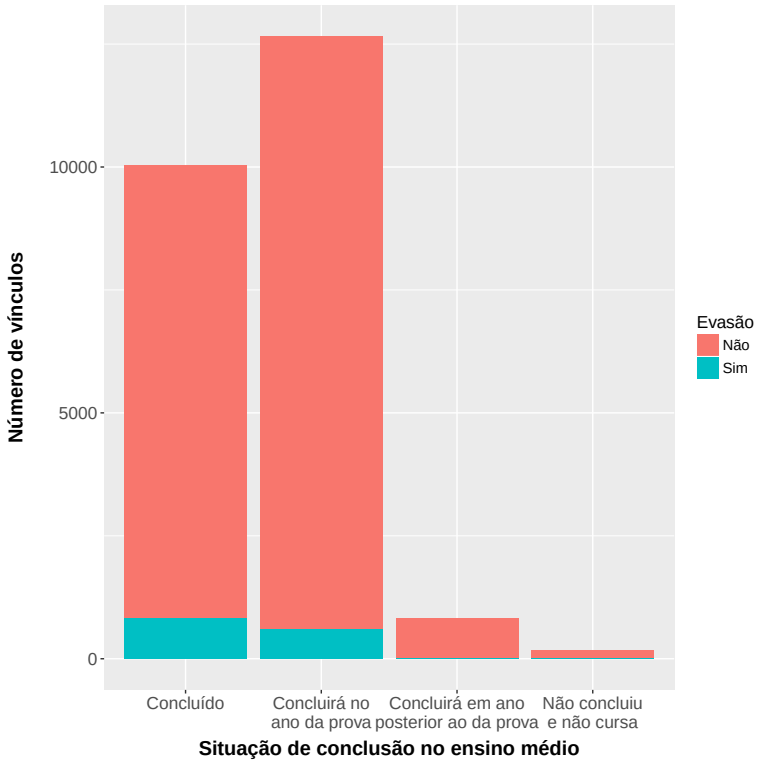

Figura 3.13: Quantidade de vínculos de alunos por evasão e situação de conclusão do Ensino Médio.

Com o objetivo de treinar os algoritmos e posteriormente obter uma estimativa dos desempenhos dos classificadores e quão bem eles generalizam, foram utilizadas a combinação de três técnicas de modelagem. Primeiro, realizou-se a separação dos dados em um grupo de *treinamento* e um de *teste*, após, foram treinados os algoritmos no grupo de *treinamento* e, em seguida, foi utilizada validação cruzada. Dessa forma, além de se obter as métricas de desempenho durante a fase de validação cruzada, que tendem a produzir estimativas otimistas sobre o desempenho em dados novos [31], os modelos classificam dados que não foram utilizados durante o treinamento. A classificação de dados não vistos previamente pelos classificadores treinados simula a obtenção de dados e, portanto, é similar a uma situação real em que novos dados são obtidos. De forma geral, é a partir da classificação de novos dados que se mede a capacidade de generalização dos classificadores treinados [31, 45].

O treinamento dos classificadores foi realizado da mesma forma nas três *tabelas* finais de evasão (nível de curso, nível de IES e nível de área de estudo). Na primeira etapa, separação das *tabelas*, elas foram divididas em duas: uma de treinamento com 75% das observações da *tabela* original e outra de teste com 25% das observações. Ambas criadas através de uma amostragem aleatória estratificada sem repetição, utilizando o atributo

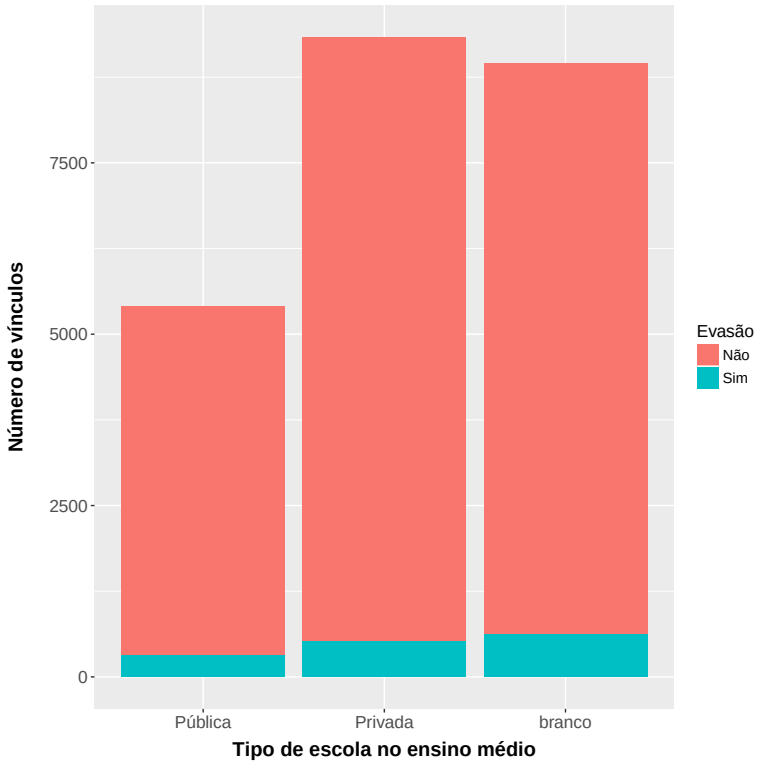

Figura 3.14: Quantidade de vínculos de alunos por evasão e tipo de escola no Ensino Médio.

"evasão" como estrato. Ou seja, as *tabelas* de treinamento e teste mantiveram a proporção da *tabela* original de vínculos rotulados como evasão. Isso pode ser visto ao comparar as Tabelas 3.12 e 3.13 a Tabela 3.11.

Em seguida, a partir de cada *tabela* de treinamento, foram criadas duas outras, uma para treinamentos com balanceamento *upsampling* e outra com balanceamento *downsampling*. Tecnicamente, qualquer *tabela* que possuir diferença entre as classes avaliadas é desbalanceada. Na literatura, é comum aparecerem desbalanceamentos da ordem de 100:1, 1.000:1 ou 10.000:1 [68]. Neste trabalho, no entanto, o maior desbalanceamento é da ordem de 18:1, no nível de evasão de IES. Apesar desse número não ser da mesma ordem do que comumente aparece na literatura, o desempenho dos modelos se mostrou superior quando utilizadas técnicas de balanceamento, que será visto na Seção 3.5.

Com o objetivo de formar novas *tabelas* que possuam a mesma quantidade de observações rotuladas evasão e retenção, os balanceamentos foram feitos a partir de amostragem com repetição das *tabelas* de treinamento. No balanceamento *upsampling*, cada rótulo deve possuir o mesmo número de observações que a quantidade do rótulo mais comum na *tabela* de treinamento [31, 45]. Por exemplo, para o nível de curso, devem existir 16.672 observações para cada rótulo. Já no *downsampling* é o contrário, o número de linhas com

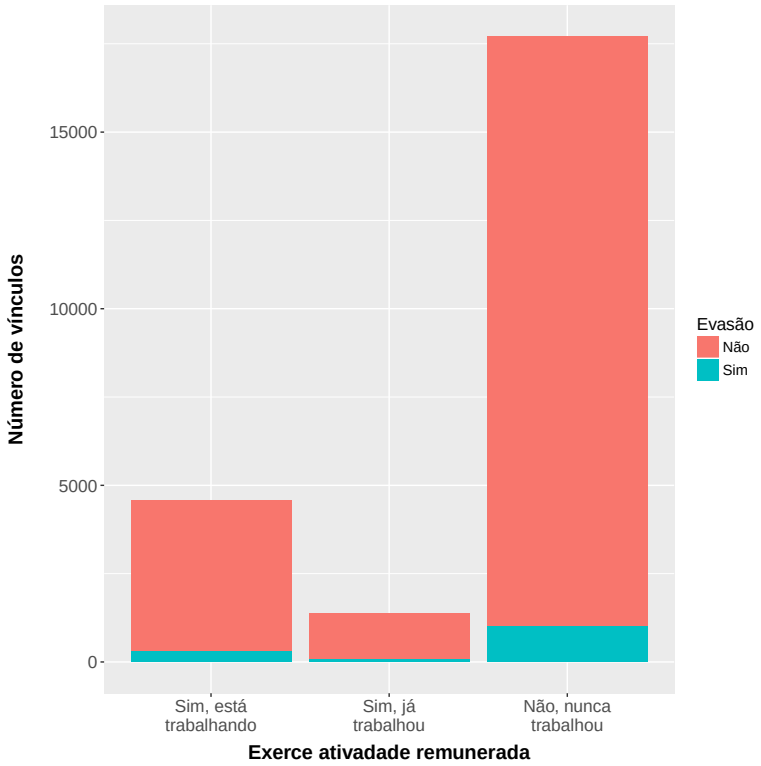

Figura 3.15: Quantidade de vínculos de alunos por evasão e situação empregatícia.

os rótulos evasão e retenção são iguais a quantidade do rótulo menos comum na *tabela* de treinamento [31, 45]. Para o caso da evasão a nível de curso, devem existir 1.098 linhas para cada rótulo. Os algoritmos de classificação foram treinados em cada uma das *tabelas* criadas, para cada nível de evasão e para cada balanceamento (*downsampling*, *upsampling* e não balanceado), utilizando validação cruzada de tamanho 4.

A validação cruzada é uma técnica que faz reamostragens para realizar o treinamento de um modelo [31, 45]. Uma validação de tamanho x realiza o procedimento de dividir a *tabela* de treinamento em x partes, em seguida realiza o treinamento do modelo em $x-1$ partes e, após, verifica o seu desempenho na parte deixada de fora. O procedimento é repetido x vezes até que todas as partes tenham servido para determinar o desempenho do modelo [31, 45]. Apesar de na literatura ser comum a utilização de validação cruzada de tamanho 10, não nada especial em relação a este número, sendo que outros valores como 4, 15 ou 20 podem ter os mesmos efeitos [31].

Por padrão, o pacote *caret* treina os algoritmos com 3^p configurações, em que p é a quantidade de parâmetros numéricos que podem assumir diferentes valores. Nos casos em que p são parâmetros booleanos, o número de configurações se modifica para 3×2^p. O valor 3 é o padrão estabelecido para a opção *tunelength*. O treinamento busca, de acordo com a métrica selecionada, a configuração que atinge o melhor desempenho

Tabela 3.12: Número de vínculos e percentual de alunos nas *tabelas* de treinamento por nível de evasão e rótulo dos vínculos.

——	Nível de evasão					
Rótulo	Curso	%	Área de estudo	%	IES	%
Evasão	1.098	6,18	942	5,94	909	5,12
Retenção	16.672	93,82	14.934	94,06	16.861	94,88
Total	**17.770**	**100,00**	**15.876**	**100,00**	**17.770**	**100,00**

Tabela 3.13: Número de vínculos e percentual de alunos nas *tabelas* de teste por nível de evasão e rótulo dos vínculos.

——	Nível de evasão					
Rótulo	Curso	%	Área de estudo	%	IES	%
Evasão	365	6,17	314	5,94	302	5,10
Retenção	5.557	93,83	4.977	94,06	5.620	94,90
Total	**5.922**	**100,00**	**5.291**	**100,00**	**5.922**	**100,00**

médio considerando os resultados da validação cruzada. Alguns algoritmos não possuem parâmetros configuráveis para se adequar melhor aos dados. O *Naive Bayes*, por exemplo, possui apenas duas configurações possíveis, enquanto a regressão logística, apenas uma. Já os outros três algoritmos, C5.0; CART e Redes Neurais, possuem um número ilimitado de possíveis configurações, já que pelo menos um dos parâmetros é numérico. No CART, apenas um parâmetro é configurável; no C5.0, existem 3 parâmetros, sendo apenas um numérico e, nas redes neurais artificias de uma camada, existem 2 parâmetros numéricos configuráveis.

Como este trabalho não possui o objetivo de encontrar um classificador ótimo para o problema, não foram feitos ajustes finos no treinamento dos algoritmos, da mesma forma que o trabalho também não focou na engenharia de atributos. Dessa forma, para os algoritmos que permitem mais de duas configurações, foram permitidas até 10 configurações diferentes (*tunelength=10*), que o próprio pacote *caret* estipula. No caso das redes neurais, foram permitidas 9 configurações, uma vez que ela possui dois parâmetros numéricos configuráveis. O objetivo foi evitar que algum algoritmo se mostrasse inadequado para a tarefa de classificação devido a escolha de determinada configuração.

A implementação do *Naive Bayes* utilizada no *caret* é a do pacote *e1071*[13]. Sua implementação assume a independência entre os atributos preditores e a classe de interesse. Como não há parâmetros configuráveis, este algoritmo foi executado em seu formato padrão. Neste formato, há duas configurações possíveis, uma que considera que a distribuição condicional a classe (evasão) é gaussiana para os atributos numéricos e, uma segunda, que utiliza uma densidade *kernel* para se aproximar aos dados de trei-

[13] `https://cran.r-project.org/web/packages/e1071/index.html`

namento. Em [69], mostrou-se que a distribuição condicional dos atributos próxima de uma distribuição gaussiana obtém desempenhos superiores à densidade *kernel*. Porém, se a distribuição dos atributos diferir significativamente de uma normal, o desempenho do classificador é superior quando utilizada a densidade *kernel* para estimar a distribuição condicional dos atributos.

A rede neural artificial utilizada foi a *feed forward multilayer perceptron* com *back propagation* de uma camada escondida. O *caret* utiliza uma implementação chamada *Stuttgart Neural Network Simulator*[14]. Como padrão, a implementação no *caret* possui 2 parâmetros numéricos configuráveis. São eles: decaimento dos pesos e o número de nós ou neurônios na camada escondida. O pacote utiliza também como padrão 3 valores distintos para cada parâmetro. O decaimento assume valores de 0; 0,0001 e 0,1; o número de nós na camada escondida recebe valores de 1, 3 e 5, formando assim, 9 configurações distintas para serem treinadas. O decaimento de pesos funciona como um determinante de quão dominante será a regularização dos pesos, ou seja, o quão penalizados serão os pesos, de acordo com seu tamanho, para o cálculo dos pesos finais. Um valor de decaimento igual a zero deixa o modelo livre para calcular seus pesos, não importando o quão grande eles sejam. Além dos parâmetros configuráveis, foi necessário alterar duas opções do modelo, a quantidade máxima de iterações permitidas até a convergência do modelo para 1.000 e a quantidade de pesos calculáveis para 3.000. Os valores padrões faziam com que os modelos definidos com mais de 1 nó na camada escondida não fossem executados. Dessa forma, foi preciso também relaxar os valores iniciais para que todas as 9 configurações do modelo fossem executadas sem paradas forçadas, isto é, anteriores ao seu critério de parada padrão. De acordo com o autor do pacote, o padrão foi estabelecido para que não fossem treinadas redes muito demoradas.

A regressão logística utilizada neste trabalho foi a implementada na função *glm* do pacote *stats*[15]. A única definição que o usuário deve escolher é a transformação utilizada para garantir que a função produza uma probabilidade. As funções mais utilizadas são a *probit* e a *logit*, que produzem resultados próximos [44]. Neste trabalho foi escolhida a *logit*. Apenas uma opção foi modificada em relação ao padrão estabelecido pelo *caret*. Assim como no treinamento das redes neurais, o número de iterações estabelecido pelo *caret* é baixo e não permitia, em algumas ocasiões, que o algoritmo executasse até o seu ponto de parada original (estabelecido pelo pacote *stats*). Dessa forma, ele foi modificado para 100, valor suficiente para que o processamento não fosse interrompido precipitadamente.

A implementação do CART utilizada no pacote *caret* é a chamada *rpart*[16], baseada no livro de Breiman *et al.* [51]. O algoritmo CART foi treinado de acordo com o padrão

[14] `https://cran.r-project.org/web/packages/RSNNS/`
[15] `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html`
[16] `https://cran.r-project.org/web/packages/rpart/index.html`

do pacote *caret*, exceto pela opção *tunelength* = 10. O único parâmetro configurável é o chamado *Complexity Parameter* (CP). Ele determina a porcentagem em que o erro de classificação diminuirá ao se fazer determinada divisão na árvore. Esse parâmetro também estabelece o critério de parada do algoritmo. No *caret*, o valor inicial de CP é zero, ou seja, a árvore inicial faz divisões até não haver mais ganho de informação e, em seguida, se faz uma amostragem com repetição de tamanho 10 dos valores de CP gerados a partir dessas divisões. Os valores de CP amostrados são então utilizados para treinar 10 árvores de decisão e, por fim, o programa escolhe a árvore que possui o melhor desempenho médio na validação cruzada baseado na métrica escolhida.

A implementação do C5.0 no pacote *caret* utiliza o C50[17], baseado no livro de Quinlan [46] e no código fonte do C5.0 implementado em linguagem C por Ross Quinlan. Os parâmetros do C5.0 no *caret* são três: tipo de modelo (árvores ou regras de decisão), retirada de atributos preditivos (verdadeiro ou falso) e número de *trials* (quantidade de modelos que devem ser utilizados para *boosting*).

O parâmetro "tipo do modelo" é fixado para que sejam treinadas apenas árvores de decisão, enquanto os outros dois são variados de acordo com o padrão do pacote *caret*. O parâmetro "retirada de atributos preditivos" varia entre verdadeiro e falso para cada valor do parâmetro de *boosting* (*trials*). Os valores que o pacote *caret* utiliza para o parâmetro *trials* são: 1, 10, 20, 30 e 40, em que 1, indica um treinamento normal, isto é, sem *boosting*. São formadas 10 configurações diferentes para o algoritmo treinar.

Nesta etapa de modelagem foi explicado o *design* dos experimentos e como cada modelo foi treinado.O *design* fez uso de três técnicas para o treinamento e aferição de desempenho dos modelos: separação em *tabelas* de treinamento e teste; balanceamento das *tabelas* com *upsampling* e *downsampling* e o uso de validação cruzada durante os treinamentos. Além disso, para cada algoritmo escolhido foram detalhadas as opções de configuração determinadas para o treinamento.

3.5 Avaliação

Nesta etapa, os modelos treinados foram avaliados por seus desempenhos em distinguir a classe de estudo, evasão e também por um prisma de objetivo do negócio. Uma revisão dos passos feitos até aqui se fez necessária, buscando determinar se alguma questão importante foi ignorada. Ao fim desta etapa, decidiu-se sobre o uso dos resultados gerados na mineração de dados.

- avaliação dos resultados;

[17] `https://cran.r-project.org/web/packages/C50/index.html`

- interpretação dos modelos gerados;
- revisão do processo e
- determinação da próxima etapa.

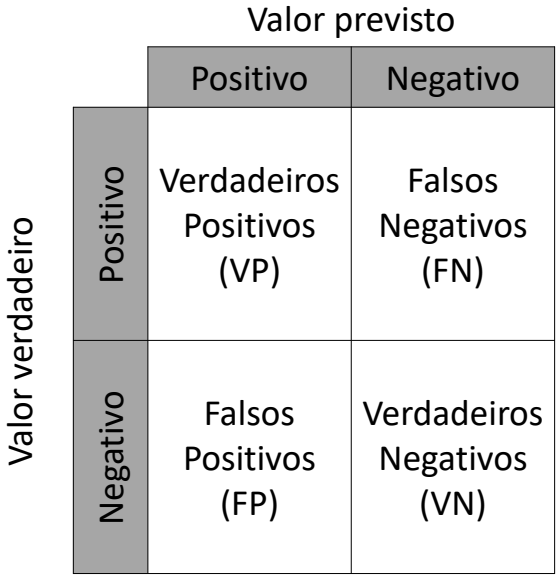

Figura 3.16: Matriz de confusão.

Para avaliar e determinar qual modelo obteve o melhor desempenho dentre os algoritmos escolhidos, foi necessária a definição de uma métrica pela qual eles seriam avaliados. A partir da matriz de confusão, Figura 3.16, é possível calcular diversas métricas de interesse. A matriz exibe todas as possibilidades de erros e acertos quando é realizada uma classificação de uma classe dicotômica, que é o caso da variável evasão. Ao definir a evasão como o caso positivo e a retenção, ou não evasão, como o caso negativo, pode-se definir o cálculo de diferentes métricas, entre elas: acurácia, especificidade e sensibilidade.

A acurácia mede o quanto um determinado classificador acerta, sejam acertos de valores positivos ou negativos. Por exemplo, se um classificador determinar, em uma *tabela* de teste, que todos as observações avaliadas são de retenção, ele obterá uma taxa de acerto superior a 90%, no entanto, ele terá errado ao classificar todos os alunos evadidos. Seu cálculo se dá pela soma dos acertos divida por todos os casos observados [31]:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.3}$$

A especificidade mede a capacidade de um modelo de classificar corretamente os valores negativos, ou seja, os alunos que não evadiram. Exemplificando, um modelo que classificar todas as observações como retenção, irá obter uma taxa de acerto de 100% nesta métrica. Seu cálculo se dá pelos acertos de valores negativos, dividido pelo total de valores negativos observados [31]:

$$ESP = \frac{VN}{VN + FP} \tag{3.4}$$

A sensibilidade é a medida complementar da especificidade e, portanto, mede a capacidade de um modelo de classificar corretamente os valores positivos, neste caso os alunos evadidos. Por exemplo, um modelo que classificar todas as observações como evasão, irá obter uma taxa de 100% de acerto nesta métrica e um valor menor que 10% na acurácia, pois assim é a distribuição da classe evasão nas *tabelas* de teste. Seu cálculo se dá pelos acertos de valores positivos, dividido pelo total de valores positivos observados [31]:

$$SEN = \frac{VP}{VP + FN} \tag{3.5}$$

A métrica escolhida para avaliar os modelos neste trabalho foi a sensibilidade. A escolha foi baseada na observação de alguns autores, [20, 25, 28], que defendem que classificar um determinado aluno erroneamente como retenção é mais prejudicial que classificá-lo erroneamente como evasão. A ideia vem da possibilidade de se usar um modelo treinado para apontar quais alunos seriam classificados como evadidos em uma nova base de dados, criando assim a oportunidade de se prover assistência aos alunos que correm risco de evadir. Os autores citados afirmam que o custo de ter um aluno em risco e não poder assisti-lo é maior que o de prestar assistência a um aluno que não corre riscos. Caso a premissa não seja verdadeira, e o custo de assistir alunos seja muito alto, seria necessário repensar a métrica utilizada neste trabalho.

Apesar de utilizar a sensibilidade como métrica principal na avaliação dos modelos gerados, a avaliação do desempenho de um modelo é um processo que não pode ser resumido a apenas um valor [31]. Dessa forma, a acurácia e a especificidade também possuem participação fundamental na avaliação da qualidade dos modelos.

Conforme detalhado na seção anterior, foram avaliados modelos para cada um dos três níveis de evasão. Cada nível de evasão foi dividido em outras três *tabelas* de treinamento, cada uma com um balanceamento diferente. Cinco algoritmos diferentes foram treinados, portanto há 45 modelos no total. A apresentação dos resultados, a seguir, foi agrupada por nível de evasão e tipo de algoritmo. Primeiro, serão detalhados os resultados para o nível de evasão de curso, em seguida da área de estudo e, por último, a nível de IES. Dentro de cada nível de evasão, a ordem de apresentação dos algoritmos é a mesma.

Tabela 3.14: Evasão a nível de curso – Desempenho médio dos modelos de *Naive Bayes* nas *tabelas* de treinamento obtido através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

——	Kernel	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	Não	78,53%	42,54%	80,91%
	Sim	91,10%	13,12%	96,24%
Downsampling	Não	62,62%	61,11%	64,12%
	Sim	66,35%	63,48%	61,39%
Upsampling	Não	64,50%	59,94%	67,80%
	Sim	65,57%	63,87%	71,62%

3.5.1 Evasão a nível de curso

Na primeira avaliação de desempenho dos algoritmos, foram realizadas análises detalhadas de cada modelo treinado, incluindo seus desempenhos durante o treinamento, a explicação de como é calculada a importância dos atributos para as classificações e a interpretação dos atributos mais importantes no contexto de negócio.

Modelos de *Naive Bayes*

Começando a avaliação dos modelos gerados através do *Naive Bayes*, tem-se a Tabela 3.14, em que são mostrados os desempenhos médios das quatro validações cruzadas realizadas durante o treinamento por tipo de balanceamento e configuração. Observando a métrica de sensibilidade, verifica-se que os modelos treinados sem qualquer balanceamento da *tabela* de treinamento obtiveram os piores desempenhos. Apesar de a acurácia dos modelos sem balanceamento serem maiores que a dos modelos balanceados, as medidas de sensibilidade e especificidade mostram que o modelo tendeu a classificar as observações como retenção, acertando menos de 50% das observações que são rotuladas como evasão. Para cada balanceamento, o *caret* seleciona o modelo com maior sensibilidade como sendo o modelo de maior desempenho nesta fase. Portanto, para os modelos treinados com balanceamento, foram selecionados os modelos com uso da função *kernel*.

A Tabela 3.15 mostra o desempenho dos melhores modelos em relação a sensibilidade ao classificar os dados de teste, que não fizeram parte de seus treinamentos. Pode-se notar que os valores são próximos dos de treinamento com validação cruzada, o que é um indicativo de que o real desempenho dos modelos pode estar próximo dos valores apresentados em ambas as tabelas. A Tabela 3.16 é a matriz de confusão do melhor modelo observado na Tabela 3.15, de onde foram calculadas as métricas apresentadas.

A Figura 3.17 mostra os intervalos de confiança para a sensibilidade dos modelos treinados ao classificar os dados de teste. Os intervalos de confiança foram calculados seguindo os princípios estatísticos de um processo de *Bernoulli* com 95% de confiança.

Tabela 3.15: Evasão a nível de curso – Desempenho dos modelos de *Naive Bayes* na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	79,26%	41,64%	81,73%
Downsampling	61,82%	66,30%	61,53%
Upsampling	**67,95%**	**61,37%**	**68,38%**

Tabela 3.16: Evasão a nível de curso – Matriz de confusão do modelo de Naive Bayes com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	224	1757
Não	141	3800

Os detalhes de como calcular esses intervalos podem ser encontrados tanto em livros de estatística [70], como em livros de mineração de dados [45].

Comparando os intervalos dois a dois, percebe-se que ambos os modelos com balanceamento obtiveram desempenhos estatisticamente superiores ao do modelo sem balanceamento. E que entre os modelos com *downsampling* e *upsampling* não houve diferença significativa, uma vez que seus intervalos se sobrepõem. Para desempatá-los, as outras métricas foram analisadas: acurácia e especificidade. Utilizando a mesma metodologia para calcular os intervalos de confiança para a sensibilidade, o modelo com *upsampling* se mostra significativamente superior em desempenho e, portanto, é considerado o melhor modelo com algoritmo *Naive Bayes* ao classificar a evasão a nível de curso.

Para esse modelo, foi analisada a importância dos atributos utilizados para construir o classificador. Para determinar a importância de cada atributo foi utilizada a curva ROC [31], que leva em consideração a sensibilidade e especificidade utilizando apenas o atributo em questão. A área abaixo da curva ROC é calculada em um plano cartesiano formado pelo eixo y (sensibilidade) e pelo eixo x (proporção de falsos positivos, calculada por $1 - especificidade$). Sendo que o ponto $(0, 1)$ indica que a classificação é perfeita, enquanto a linha diagonal representa o desempenho de uma classificação aleatória.

Na Tabela 3.17 pode-se ver quais são os atributos mais relevantes na classificação da evasão a nível de curso pelo modelo de *Naive Bayes*. As importâncias foram padronizadas para que a comparação de relevância fosse facilitada. O atributo mais importante automaticamente recebeu valor 100 e os outros atributos aparecem em termos proporcionais ao atributo mais importante. Neste caso, o atributo preditivo mais relevante foi o semestre de ingresso do aluno na UnB. O segundo mais importante foi a quantidade de vínculos com IES diferentes que o aluno possui (num_IES), seguido das notas obtidas nos exames

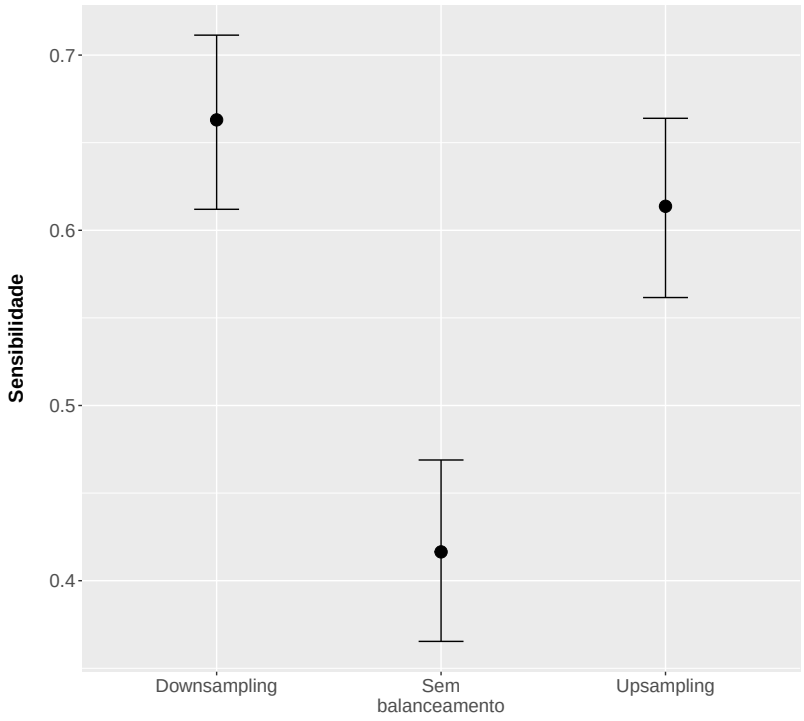

Figura 3.17: Evasão a nível de curso – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de *Naive Bayes*.

do ENEM. Dentre as notas, observa-se que a de redação foi a menos relevante e que a nota em ciências humanas foi a mais relevante. As listas completas com o detalhamento dos atributos podem ser encontradas nas pastas de dicionários na página do projeto deste trabalho[18].

O modelo obtido indica que os alunos que ingressam na universidade no primeiro semestre tendem a evadir mais. A probabilidade condicional de o aluno ser do primeiro semestre, dado que ele evadiu é de aproximadamente 80%. Já os alunos que não evadiram, a probabilidade condicional de ser do primeiro semestre é aproximadamente 50%. Para atributos numéricos, como são os casos de 'num_IES' e as notas no ENEM, o algoritmo calcula a densidade condicional a evasão desses atributos. Para o atributo que indica a quantidade de vínculos com diferentes IES, o modelo percebe que, alunos que possuem valores maiores que 1 tendem a evadir mais, ou seja, alunos que dividem a atenção em mais de uma IES possuem uma tendência maior de abandono do curso. Para as quatro notas das provas objetivas e para a redação a interpretação é a mesma, os alunos que tiram notas acima da média tendem a evadir mais. Esse é um resultado contraintuitivo, pois espera-se que alunos com pior desempenho tenham uma probabilidade de evasão maior.

[18] `https://github.com/lucke71/Classify_dropout`

Tabela 3.17: Evasão a nível de curso – Importância de atributos para o classificador *Naive Bayes* treinado com *upsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	82,11
nota_ch	78,36
nota_cn	73,27
nota_lc	72,44
nota_mt	67,87
nu_nota_redacao	60,52
st_conclusao	55,38
idade	53,87
concluiu_ens_med	52,52
ano_concluiu	49,82
id_dependencia_adm_esc	44,53
sit_func_esc	43,37
id_localizacao_esc	43,30
mun_esc_dif_prova	42,17
mun_res_dif_esc	41,82
uf_esc	36,44
nu_nota_comp3	32,44
nu_nota_comp2	30,82
uf_nasc	27,17

Tabela 3.18: Evasão a nível de curso – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	93,82%	0,00%	100,00%
1	0,0001	93,82%	0,00%	100,00%
1	0,1000	93,60%	9,11%	99,17%
3	0,0000	93,82%	0,00%	100,00%
3	0,0001	93,82%	0,00%	100,00%
3	0,1000	91,79%	15,58%	96,81%
5	0,0000	93,82%	0,00%	100,00%
5	0,0001	93,82%	0,00%	100,00%
5	0,1000	90,72%	18,22%	95,49%

Tabela 3.19: Evasão a nível de curso – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	50,87%	88,18%	13,59%
1	0,0001	51,78%	36,50%	67,15%
1	0,1000	63,30%	59,38%	67,21%
3	0,0000	57,47%	53,64%	61,30%
3	0,0001	54,73%	72,96%	36,50%
3	0,1000	61,39%	58,65%	64,12%
5	0,0000	54,51%	66,77%	42,25%
5	0,0001	55,38%	39,25%	71,48%
5	0,1000	65,12%	66,40%	63,84%

Modelos de Redes Neurais

Analisando os resultados para as redes neurais na Tabela 3.18, percebe-se que o treinamento realizado sem balanceamento obteve resultados de sensibilidade abaixo de 20% em todas as configurações. Em quase todas as configurações, o modelo classificou todas as observações como retenção. As exceções ocorreram quando o decaimento atingiu 0,1. O comportamento dos modelos treinados com balanceamento foi diferente do modelo sem balanceamento, conforme resultados nas Tabelas 3.19 e 3.20. Nos modelos com balanceamento foram registrados alguns valores de sensibilidade acima de 85%. O melhor modelo com *downsampling* é o de configuração com apenas um nó na camada escondida e decaimento igual a zero, com sensibilidade de 88,18%. O melhor modelo com *upsampling* é o que utiliza cinco nós na camada escondida e decaimento igual a 0,1 com sensibilidade igual a 90,37%.

Tabela 3.20: Evasão a nível de curso – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	54,12%	52,75%	55,48%
1	0,0001	55,06%	38,05%	72,07%
1	0,1000	64,13%	60,89%	67,38%
3	0,0000	60,40%	68,96%	51,84%
3	0,0001	65,93%	46,89%	84,97%
3	0,1000	82,34%	81,49%	83,19%
5	0,0000	69,36%	57,33%	81,40%
5	0,0001	71,42%	68,13%	74,72%
5	0,1000	86,71%	90,37%	83,06%

Tabela 3.21: Evasão a nível de curso – Desempenhos dos modelos de redes neurais na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	90,66%	21,92%	95,18%
Downsampling	**62,17%**	**65,21%**	**61,98%**
Upsampling	81,29%	50,41%	83,32%

Tabela 3.22: Evasão a nível de curso – Matriz de confusão do modelo de Redes Neurais com *downsampling*.

——	Referência	
Predição	Sim	Não
Sim	238	2113
Não	127	3444

Comparando os resultados de desempenho nas *tabelas* de treinamento com os desempenhos ao classificar os dados de teste na Tabela 3.21, percebe-se que pode ter havido *overfitting* nos modelos balanceados, ou seja, é provável que os modelos tenham se ajustado excessivamente aos dados disponíveis em treinamento, perdendo a capacidade de classificar novos dados [31, 52]. Um indício disso é a diferença entre os desempenhos de classificação dos dados de teste em relação aos desempenhos médios de treinamento. No entanto, os desempenhos das redes neurais treinadas em *tabelas* balanceadas obtiveram resultados similares aos do *Naive Bayes*. Em comparação com os resultados de desempenho do *Naive Bayes*, nos modelos de redes neurais há diferença significativa entre os modelos treinados com *downsampling* e *upsampling*, conforme mostra a Figura 3.18. Indicando que o modelo com *downsampling* é o melhor modelo de redes neurais para evasão a nível de curso.

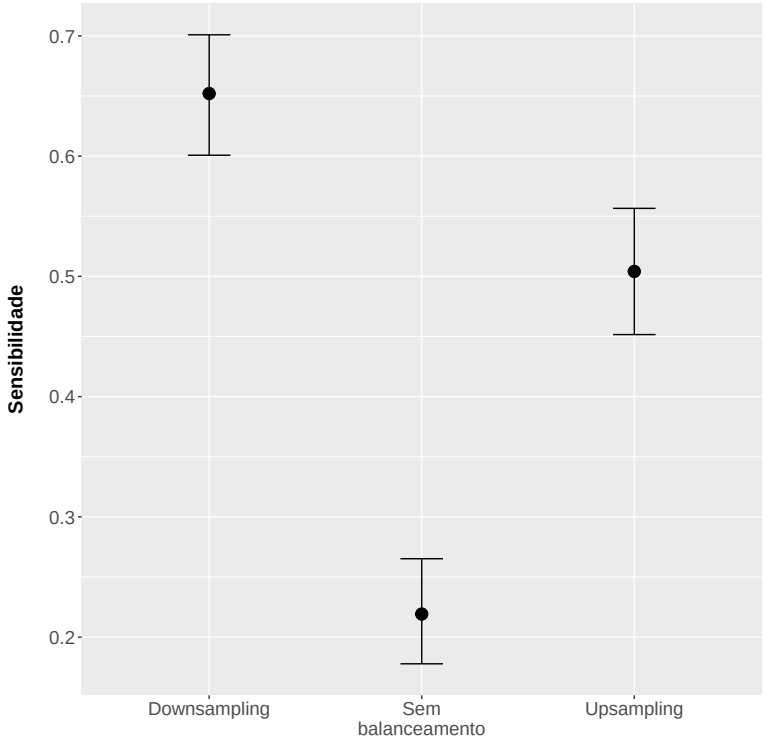

Figura 3.18: Evasão a nível de curso – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de redes neurais.

Baseado no trabalho de Garson [71], é possível identificar o grau de importância de cada atributo de um modelo de rede neural, através do tamanho dos pesos calculados entre as camadas de entrada, em que ficam os atributos, e a escondida, em que ficam os nós ou neurônios. No entanto, esta técnica não permite identificar se o atributo em questão é um fator positivo ou não para a classe estudada.

Tabela 3.23: Evasão a nível de curso – Importância de atributos para o classificador de redes neurais treinado com *downsampling*.

Nome do atributo	Importância
ano_concluiu	100,00
nu_nota_comp2	46,10
doc_exercicio	38,29
nu_nota_comp5	37,97
vagas	36,47
nu_nota_comp3	31,85
uf_prova21	29,77
q035G	29,55
uf_res28	29,55
uf_prova17	29,54
q00420	29,53
mun_res_dif_escTRUE	29,51
q001G	29,49
uf_res14	29,39
q0253	29,19
q0283	29,17
q039B	29,15
q019B	29,13
q0465	29,10
q001E	29,05

Observando a Tabela 3.23, percebe-se que o modelo de redes neurais não considerou como importantes os mesmos atributos que o modelo de *Naive Bayes*. As importâncias dos atributos também foram padronizadas de forma que o atributo com maior peso recebeu valor 100 e os outros atributos apareceram em termos proporcionais a ele. O atributo mais importante para o modelo de rede neural foi o ano em que o aluno concluiu o ensino médio, que também apareceu no modelo de *Naive Bayes*, porém com menos importância. Além disso, observa-se que as notas dos exames do ENEM, neste modelo, deram lugar às notas de competência, que são extraídas da capacidade do aluno de escrever a redação. Outra diferença foi que, para os atributos não numéricos, foi necessário observar o valor que eles assumem. Um exemplo é a questão 035 (q035) do questionário do ENEM. Ela aparece na tabela com um valor, 'G', preenchido. A questão 035 pergunta ao aluno sobre o tipo de escola que ele cursou o ensino médio. A opção 'G' indica que o aluno estudou apenas em escola situada em comunidade quilombola.

Modelos de Regressão logística

Na Tabela 3.24, pode-se ver o desempenho médio medido na base de treinamento para os modelos de regressão logística. Percebe-se que, mais uma vez o modelo sem balanceamento

Tabela 3.24: Evasão a nível de curso – Desempenho médio dos modelos de regressão logística nas *tabelas* de treinamento obtidos através da validação cruzada tamanho 4 por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,51%	9,57%	99,03%
Downsampling	67,08%	68,49%	65,66%
Upsampling	77,28%	78,78%	75,78%

Tabela 3.25: Evasão a nível de curso – Desempenho dos modelos de regressão logística na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,62%	8,22%	99,23%
Downsampling	68,07%	69,59%	67,97%
Upsampling	**74,15%**	**59,73%**	**75,09%**

Tabela 3.26: Evasão a nível de curso – Matriz de confusão do modelo de Regressão Logística com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	218	1384
Não	147	4173

classificou grande parte das observações como retenção, obtendo sensibilidade inferior a 10%. Os modelos treinados com balanceamento obtiveram novamente desempenho superior, indicando que o balanceamento da *tabela* de treinamento é importante para obter um modelo capaz de identificar os alunos evadidos.

Na Tabela 3.25, pode-se ver que o desempenho do modelo treinado sem balanceamento continua acertando apenas a classificação dos alunos não evadidos. Nota-se também que, o desempenho do modelo treinado com *downsampling* é bastante similar ao desempenho visto na validação cruzada, enquanto o desempenho do modelo com *upsampling* não. Isso indica que o modelo com *upsampling* sofre de *overfitting*. O resultado na Tabela 3.25 indica que o modelo com *downsampling* é, assim como com as redes neurais, o modelo com melhor desempenho. No entanto, a 95% de confiança, a diferença entre os modelos com balanceamento não é estatisticamente significante, conforme Figura 3.19. Seguindo o procedimento para os modelos de *Naive Bayes*, o desempate é realizado com uma análise das outras métricas. Assim como no caso do *Naive Bayes*, o modelo com *upsampling* se mostra estatisticamente superior, nas outras métricas, ao modelo com *downsampling*. A Tabela 3.26 traz a matriz de confusão para o modelo de regressão logística com *downsampling*.

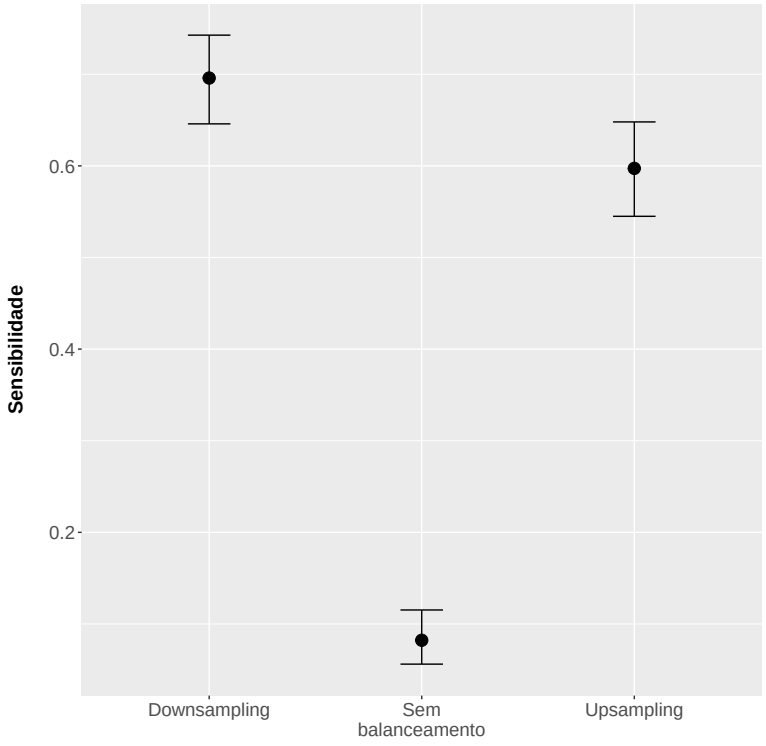

Figura 3.19: Evasão a nível de curso – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de regressão logística.

Tabela 3.27: Evasão a nível de curso – Importância de atributos para o classificador de regressão logística treinado com *upsampling*.

Nome do atributo	Importância
semestre_ingresso2	100,00
num_ies	78,03
ano_ing	53,17
in_compl_monitoria1	26,06
nota_mt	21,98
q0262	20,50
candidatos	18,98
ano_ces2012	18,42
q0265	17,60
nota_cn	17,32
doc_integ_sem_de	16,40
q0244	15,75
ano_ces2011	14,92
ano_ces2013	14,30
ano_enem2011	14,24
q007C	14,13
nota_ch	13,81
q0264	13,53
nu_nota_redacao	13,48
q041B	12,99

Para determinar a importância dos atributos da regressão logística foi utilizado o coeficiente estimado para cada atributo, quanto maior o coeficiente estimado, maior a importância. Assim como nas redes neurais, atributos não numéricos são tratados com a criação de atributos para cada valor que o atributo pode assumir, também conhecidos como variáveis *dummy*. Observando a Tabela 3.27, nota-se que os três atributos considerados mais importantes foram o fato do aluno ter ingressado no segundo semestre, a quantidade de vínculos a IES distintas a que ele possui e o ano em que o aluno ingressou na universidade.

O modelo indica que alunos que ingressam no segundo semestre, têm chances 7 vezes menor de evadir se comparada aos que ingressam no primeiro semestre. Em relação ao número de IES distintas, o modelo indica uma chance 3 vezes maior de evadir para o aumento de cada unidade nesse atributo. Ou seja, um aluno que está simultaneamente em duas IES possui uma chance de evadir 3 vezes maior que um aluno que está apenas em uma. É importante ressaltar que nas análises de razões de chances de um determinado atributo, supõe-se que os demais atributos mantenham-se constantes.

Modelos CART

Os resultados médios dos modelos CART durante a validação cruzada foram trazidos em três figuras. A Figura 3.20 mostra resultados semelhantes aos já vistos nos modelos dos algoritmos analisados anteriormente. O modelo treinado sem balanceamento acaba por classificar grande parte das observações como retenção e obtém um valor de sensibilidade inferior aos dos modelos treinados em *tabelas* balanceadas. As Figuras 3.21 e 3.22 mostram os modelos treinados com *downsampling* e *upsampling*, respectivamente. Em ambas, percebe-se que o parâmetro de complexidade que atinge o maior valor de sensibilidade se encontra entre 0,05 e 0,1. O comportamento dos modelos balanceados são parecidos nas diferentes configurações. A única diferença foi no comportamento da especificidade quando o parâmetro de complexidade foi superior a 0,2.

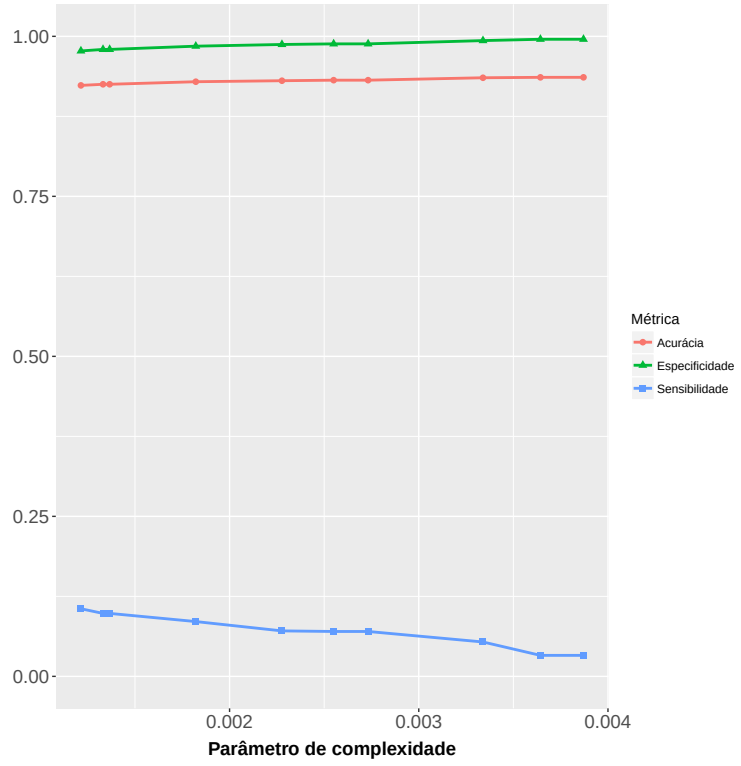

Figura 3.20: Evasão a nível de curso – Desempenho médio dos modelos CART nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

Observando a Tabela 3.28, percebe-se que os valores para sensibilidade foram superiores aos encontrados na média da validação cruzada. Em compensação, os valores das outras métricas foram inferiores. Segundo este experimento, os modelos treinados com balanceamento foram capazes de identificar os alunos rotulados como evasão da base de

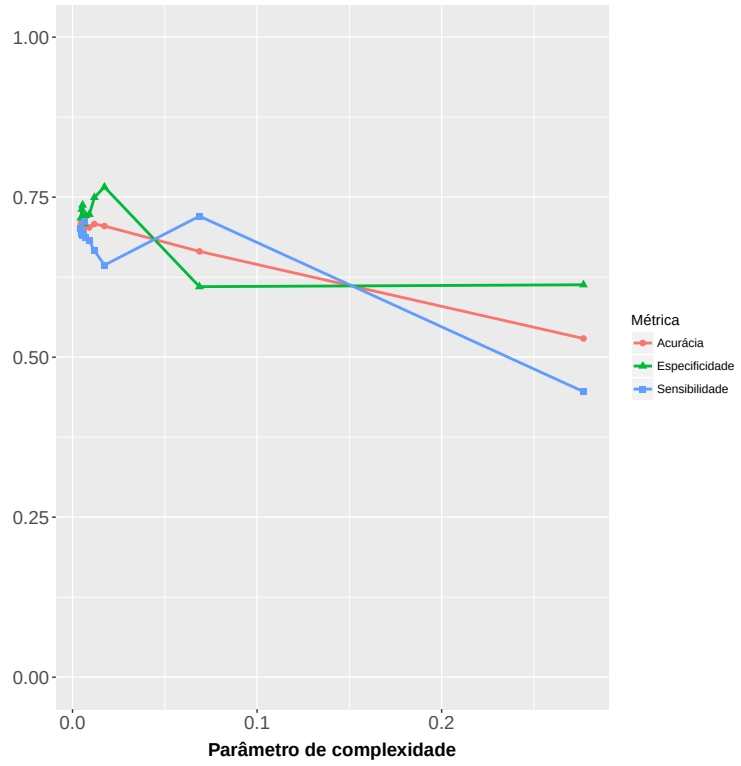

Figura 3.21: Evasão a nível de curso – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

testes, no entanto, erraram metade das classificações realizadas. Ou seja, não conseguiram identificar com a mesma qualidade os evadidos e os retidos.

A Figura 3.23, mostra que, em relação à sensibilidade, os modelos balanceados não apresentaram diferença estatisticamente significante entre eles e que ambos foram superiores quando comparados ao modelo sem balanceamento.

Para determinar a importância dos atributos dos modelos CART foi utilizado o método proposto por Breiman [51]. A cada divisão realizada na criação da árvore, foi computado o ganho de informação obtido ao dividir as observações por determinado atributo. Repetindo o procedimento anterior, o atributo com maior importância recebeu valor 100,

Tabela 3.28: Evasão a nível de curso – Desempenho dos modelos CART na *tabela* de teste por tipo de balanceamento.

———	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	92,91%	7,12%	98,54%
Downsampling	**50,22%**	**83,84%**	**48,01%**
Upsampling	50,15%	81,10%	48,12%

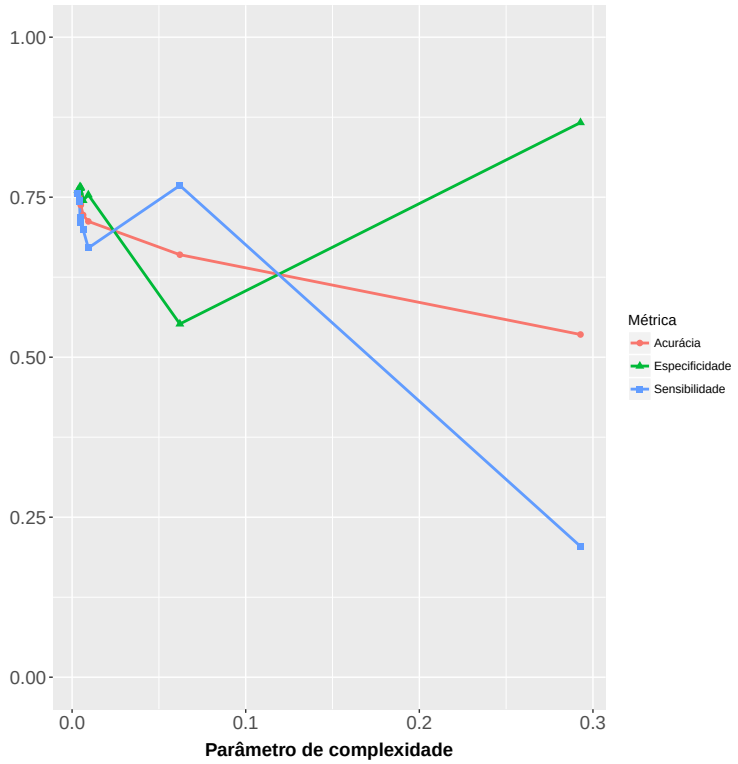

Figura 3.22: Evasão a nível de curso – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

enquanto os demais apresentam valores proporcionais ao atributo de maior relevância.

Como o CART permite que a árvore seja podada, é possível que atributos que não formam a árvore final receberem importância diferente de zero. Isso pode ocorrer quando a divisão que seria feita por esse atributo não traz ganho maior que o determinado para o treinamento do modelo. Observando a Tabela 3.33, verifica-se que apenas cinco atributos tiveram valores maiores que zero, enquanto que a árvore final foi composta por apenas um atributo, o semestre de ingresso. Assim como no *Naive Bayes* e na regressão logística, o semestre de ingresso é o atributo mais importante e sua interpretação continua a mesma, ingressantes do primeiro semestre são classificados como evasão pelo modelo CART, enquanto ingressantes do segundo semestre como retenção.

Tabela 3.29: Evasão a nível de curso – Matriz de confusão do modelo de CART com *downsampling*.

———	Referência	
Predição	Sim	Não
Sim	306	2889
Não	59	2668

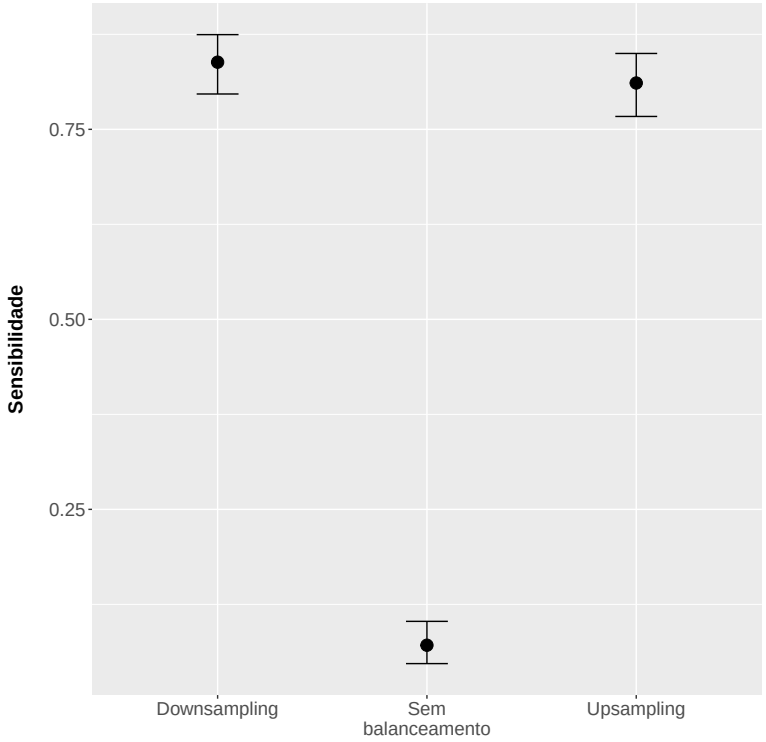

Figura 3.23: Evasão a nível de curso – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos CART.

Modelos C5.0

Os resultados médios de desempenho dos modelos C5.0 durante a validação cruzada são apresentados nas Figuras 3.24, 3.25 e 3.26. Na Figura 3.24, percebe-se que da mesma forma que ocorreu nos outros algoritmos, o treinamento sem balanceamento também classificou as observações como retenção, obtendo sensibilidade zero em todas as configurações. Os pontos que indicam se houve retirada automática de atributos durante o treinamento (triângulos ou círculos) não são claros, pois os resultados de desempenho das configurações se sobrepõem com exatidão, independentemente do número de *trials*.

Tabela 3.30: Evasão a nível de curso – Importância de atributos para o classificador CART treinado com *downsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	70,63
nota_ch	47,39
nota_cn	38,28
nota_lc	36,67

Na Figura 3.25, observa-se o desempenho do modelo treinado com *downsampling*. Esta não segue a mesma escala da Figura 3.24, mostrando apenas valores de 67% até 75%. A justificativa da utilização de uma escala diferenciada é que se a mesma escala fosse mantida não seria possível identificar visualmente as diferenças entre as configurações. Percebe-se que retirar ou não atributos automaticamente durante o treinamento apresentou comportamentos diferentes entre as métricas analisadas. Para a sensibilidade, não retirar atributos trouxe um desempenho superior para o modelo quando o número de *trials* cresce. Portanto, o melhor modelo, neste balanceamento, é o que apresenta *trials* igual a 40 e utiliza todos atributos disponíveis.

A Figura 3.26 traz o desempenho do modelo treinado com *upsampling*. Mais uma vez, a escala não é a mesma das Figuras 3.24 e 3.25 e os valores variam de 90% a 100%. O modelo treinado com *upsampling* consegue classificar corretamente praticamente todas as observações, sejam elas evasão ou retenção, o que por si só, é um indicador de *overfitting*. Como várias configurações apresentam o mesmo resultado, o modelo a apresentar o melhor resultado com o menor número de *trials* foi o escolhido como o melhor modelo deste balanceamento. A configuração é número de *trials* igual a 10 com retirada de atributos.

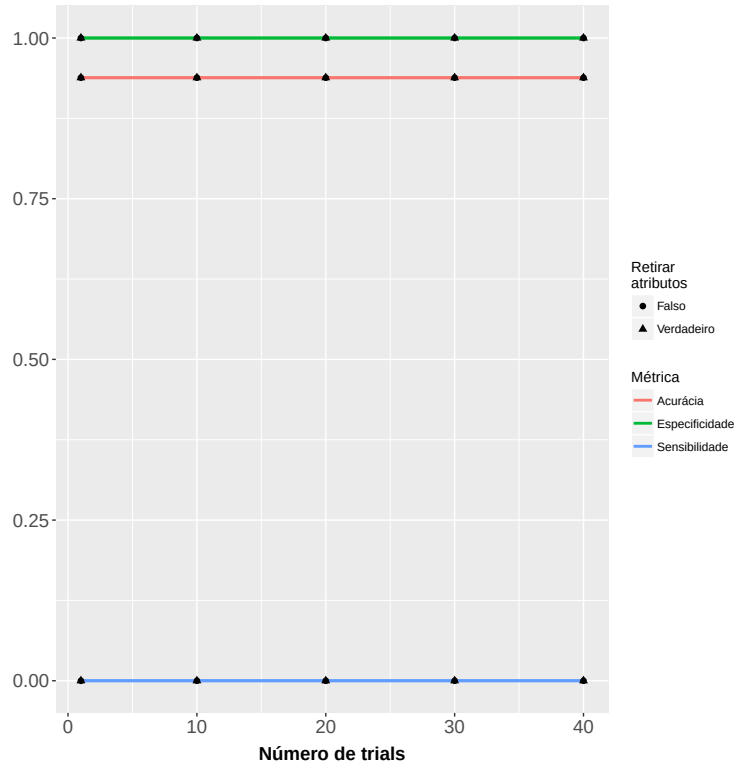

Figura 3.24: Evasão a nível de curso – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

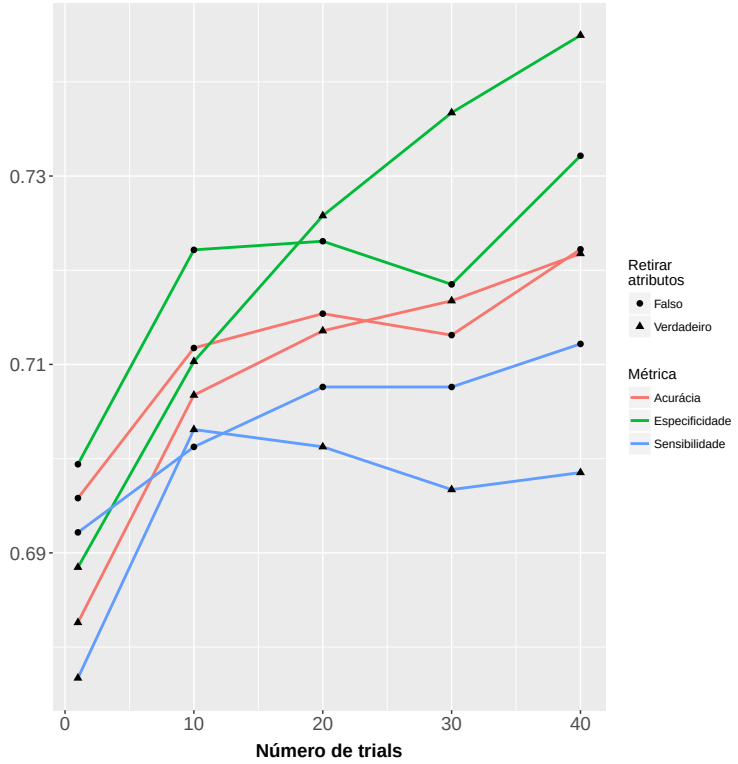

Figura 3.25: Evasão a nível de curso – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

Analisando a Tabela 3.31, nota-se que o modelo treinado sem balanceamento continua com o mesmo comportamento apresentado durante o treinamento. Além disso, a suspeita de *overfitting* no modelo treinado com *upsampling* ganha mais evidência com o resultado de sensibilidade desta tabela. Enquanto o modelo mostrava acertar a classificação de todas as instâncias durante o treinamento, ele não foi capaz de acertar mais de 10% das observações rotuladas como evasão na base de teste. O modelo com *downsampling* apresentou resultados similares aos vistos durante seu treinamento, indicando que os valores apresentados devem ser próximos aos valores reais de sua capacidade de classificação.

A Figura 3.27 mostra a diferença de desempenho, segundo a sensibilidade dos modelos treinados com diferentes formas de balanceamento. Os modelos sem balanceamento e com *upsampling* mostram não haver diferença significativa entre eles, enquanto o modelo com *downsampling* mostra que, com 95% de confiança, seu valor real de sensibilidade está entre 65% e 75%.

A forma de determinar a importância de atributos no C5.0 é muito similar a implementada no CART. A diferença principal é que o C5.0 utiliza *boosting* e, portanto, a importância dos atributos é somada para cada árvore utilizada. Dessa forma, é possível

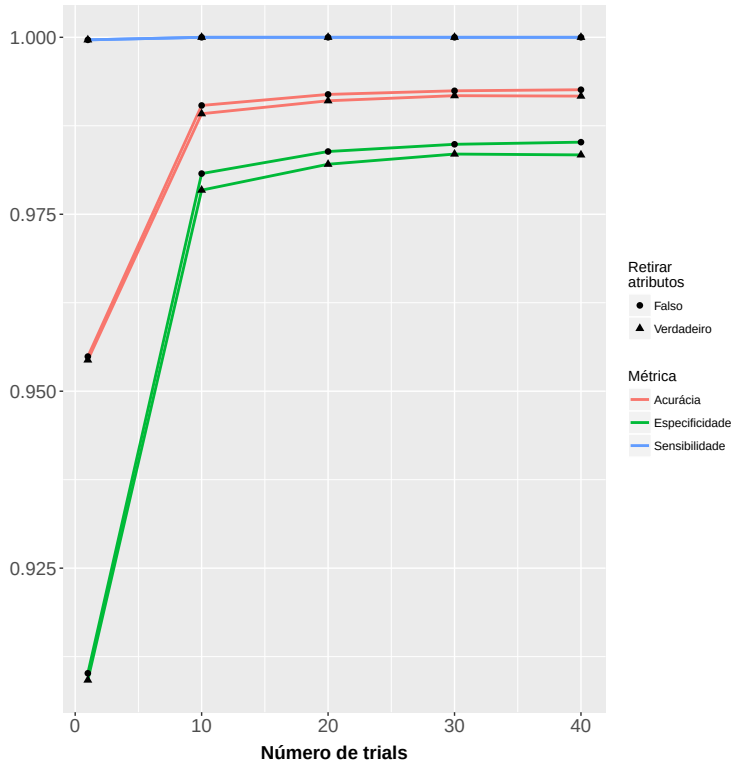

Figura 3.26: Evasão a nível de curso – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

haver mais de um atributo com importância igual a 100.

Observando a Tabela 3.33, verifica-se que existem 11 atributos com importância igual a 100, o que significa que esses atributos foram utilizados como a primeira divisão em pelo menos uma das árvores treinadas e utilizadas no modelo final. Mais uma vez o semestre de ingresso apareceu como atributo importante, assim como o número de IES distintas a que o aluno possui vínculo. A interpretação desses atributos e dos outros já vistos em modelos passados, como as notas no ENEM não é alterada. Alguns atributos só foram considerados importantes pelo C5.0, como é o caso de atividades complementares dos alunos: estágio, pesquisa e monitoria; uso de cadeiras de rodas; tempo, em anos, para

Tabela 3.31: Evasão a nível de curso – Desempenho dos modelos C5.0 na *tabela* de teste por tipo de balanceamento.

———	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,68%	3,29%	99,62%
Downsampling	**74,55%**	**69,59%**	**74,88%**
Upsampling	93,03%	8,49%	98,58%

Tabela 3.32: Evasão a nível de curso – Matriz de confusão do modelo de C5.0 com *downsampling*.

———	Referência	
Predição	Sim	Não
Sim	254	1396
Não	111	4161

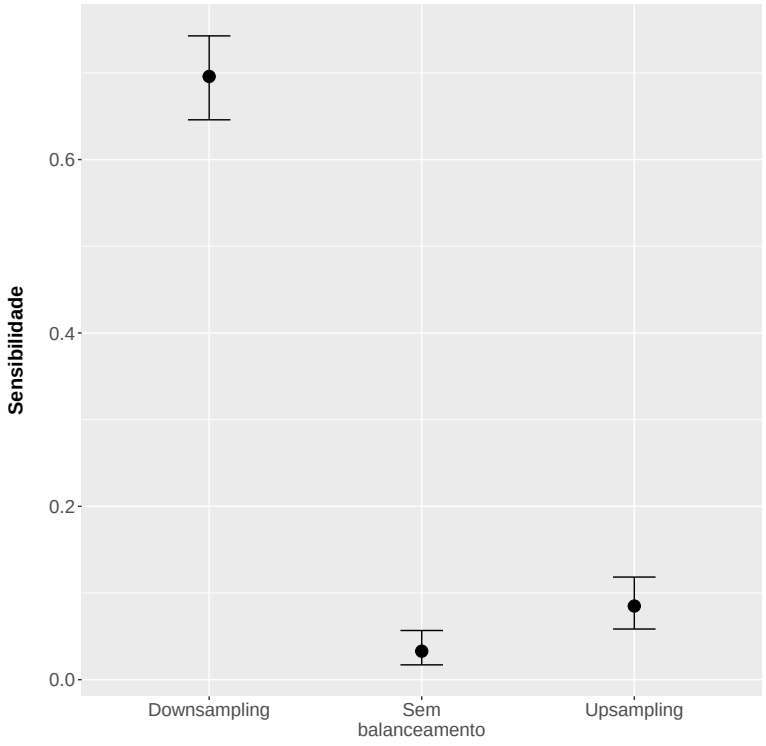

Figura 3.27: Evasão a nível de curso – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos C5.0.

integralização no curso de turno integral e a situação do aluno quanto a conclusão do ensino médio no momento de realização do ENEM.

O modelo indica que alunos que fazem a atividade de estágio e pesquisa são mais propensos a evadir, enquanto os que fazem a atividade de monitoria são menos propensos. O fato do aluno utilizar cadeira de rodas não parece atingir diretamente muitos nós, apesar de ter sido utilizado como primeiro divisor em uma das árvores. O modelo aponta apenas três alunos com cadeira de rodas como evasão enquanto o restante dos alunos são divididos por outros atributos. Já a questão sobre a situação do aluno quanto à conclusão no ensino médio aponta que alunos que estão cursando o ensino médio e se formarão após o ano em que realizaram a prova do ENEM (treineiros ou alunos do último ano) tendem a não evadir.

Tabela 3.33: Evasão a nível de curso – Importância de atributos para o classificador CART treinado com *downsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	100,00
ano_ing	100,00
nota_cn	100,00
nota_ch	100,00
in_mesa_cadeira_rodas	100,00
nu_integralizacao_integral	100,00
in_compl_estagio	100,00
in_compl_pesquisa	100,00
st_conclusao	100,00
in_compl_monitoria	100,00
uf_prova	99,95
q034	99,64
ano_ces	99,18
q030	98,91
uf_nasc	98,77
nota_mt	97,13
q026	96,77
ano_enem	94,63
uf_esc	89,75

Comparação entre os algoritmos

Depois de analisados todos os modelos para evasão a nível de curso, fez-se necessária uma comparação entre os algoritmos utilizados. Para realizar essa comparação, foram escolhidos os modelos com melhor desempenho, baseado na sensibilidade ao classificar as observações da *tabela* de teste. Nos casos de empate, ou seja, em que a diferença não foi estatisticamente significativa, escolheu-se os modelos com melhor desempenho nas outras métricas. Para os modelos que empataram em todas as métricas, selecionou-se o modelo de *downsampling*, pois ele oferece um custo computacional menor ao ser treinado. Dessa forma, *Naive Bayes* e regressão logística tiveram como melhor modelo o treinado com *upsampling*, para todos os outros algoritmos foram escolhidos os modelos treinados com *downsampling*.

A Figura 3.28 mostra a comparação entre esses modelos. Foram computados intervalos de confiança com correção de Bonferroni para 95% de confiança. Ao realizar comparações entre vários grupos, e avaliar intervalos de confiança dois a dois múltiplas vezes, aumenta-se a probabilidade de que uma das comparações esteja errada [70]. A probabilidade de se cometer um erro ao comparar cinco grupos dois a dois com intervalos de confiança de 95% para cada grupo é de aproximadamente 30%. A correção de Bonferroni deixa os intervalos mais conservadores, modificando o nível de confiança para cada intervalo calculado, de forma que ao realizar todas as comparações possíveis, o nível final de confiança seja 95%. Como são cinco grupos, existem dez comparações possíveis dois a dois, portanto, divide-se o nível de significância (5%) por dez, obtendo uma confiança de 99,5% para cada intervalo.

Analisando os resultados da Figura 3.28, pode-se notar que apenas um modelo se sobressaiu sobre os outros: o treinado a partir do algoritmo CART. Dessa forma, baseado na métrica escolhida, há evidências estatísticas de que o CART seria o algoritmo, dentre os comparados, com o melhor desempenho para evasão a nível de curso.

3.5.2 Evasão a nível de área de estudo

Neste nível de evasão, as tabelas e gráficos que mostram os desempenhos dos modelos durante a fase de treinamento, isto é, baseado nos valores médios obtidos nas validações cruzadas, foram apresentados no Apêndice A. O foco das análises foi, então, no desempenho dos modelos ao tentar classificar os dados de teste e nas análises interpretativas de como os modelos realizam suas classificações em evasão e retenção.

Modelos de *Naive Bayes*

Seguindo a mesma ordem das análises a nível de curso, a primeira é novamente sobre os modelos gerados a partir do *Naive Bayes*. A Tabela 3.34 mostra resultados muito simila-

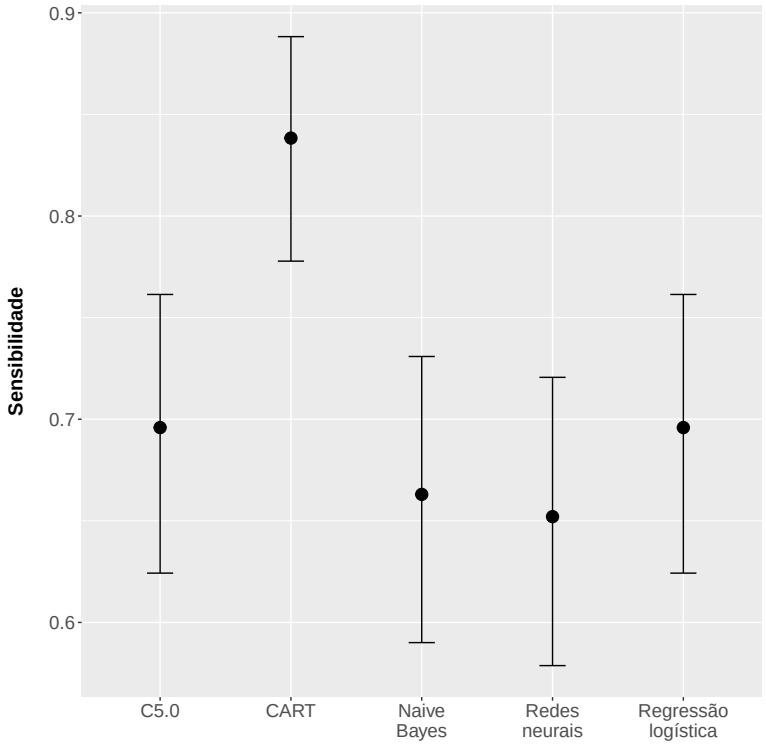

Figura 3.28: Evasão a nível de curso – Intervalos de confiança com correção de Bonferroni a 95% para a sensibilidade medida na *tabela* de teste para os melhores modelos de cada algoritmo.

res aos encontrados na avaliação da evasão a nível de curso. Mais uma vez, baseado em sensibilidade, o modelo treinado com *downsampling* obteve o melhor resultado quando classificou os dados de teste. No entanto, observando a Figura 3.29, essa diferença não foi estatisticamente significante a 95% de confiança. Além disso, ambos os modelos balanceados obtiveram desempenhos superiores, em termos de sensibilidade, em relação ao modelo treinado sem balanceamento. O modelo com melhor desempenho neste grupo de análise foi o treinado com *downsampling*. Não há diferença significativa entre ele e o modelo treinado com *upsampling*, e portanto, o modelo que possui o menor custo computacional foi escolhido. A configuração do modelo treinado com *downsampling* escolhida durante o treinamento foi a que não utiliza a função *kernel* para estimar a densidade dos atributos numéricos, ou seja, foi utilizada uma distribuição gaussiana.

Verificando a Tabela 3.36, percebe-se que os resultados são semelhantes aos vistos na avaliação da evasão a nível de curso para este algoritmo. Os dois atributos mais importantes foram os mesmos e aparecem na mesma ordem. Em seguida, as quatro notas dos exames objetivos, embora não estejam na mesma ordem. Há uma mudança também em relação ao nível de importância dado para cada atributo. É importante lembrar que o

Tabela 3.34: Evasão a nível de área de estudo – Desempenho dos modelos de *Naive Bayes* na *tabela* de teste por tipo de balanceamento.

———	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	78,44%	42,68%	80,69%
Downsampling	**65,98%**	**62,42%**	**66,20%**
Upsampling	67,30%	56,37%	67,99%

Tabela 3.35: Evasão a nível de área de estudo – Matriz de confusão do modelo de Naive Bayes com *downsampling*.

———	Referência	
Predição	Sim	Não
Sim	196	1682
Não	118	3295

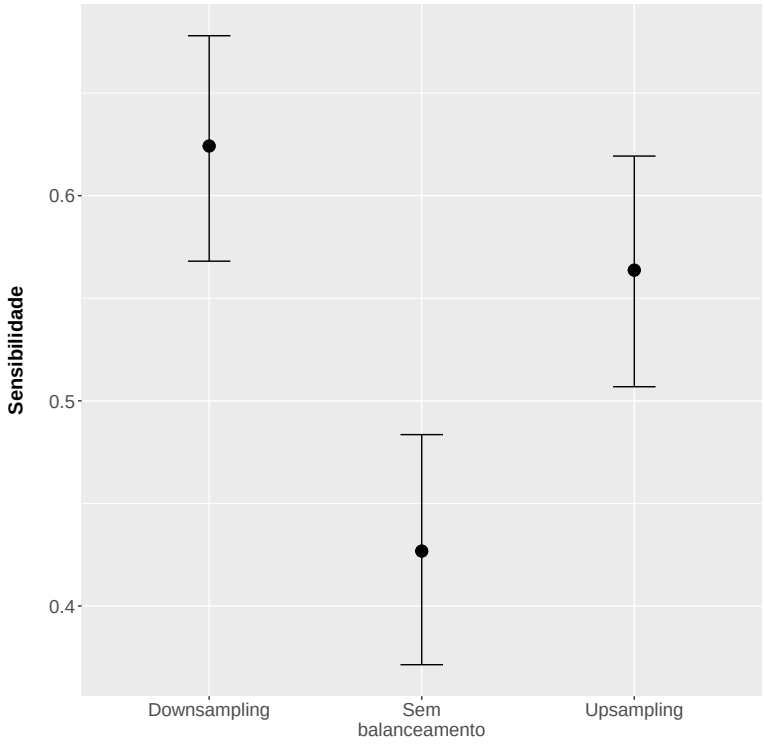

Figura 3.29: Evasão a nível de área de estudo – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de *Naive Bayes*.

modelo avaliado na evasão a nível de curso foi o treinado com *upsampling*, enquanto que a nível de área, o avaliado é o treinado com *downsampling*.

Apesar de os modelos terem sido treinados com balanceamentos distintos e analisarem outro nível de evasão, as interpretações dos atributos não são muito diferentes. Os

Tabela 3.36: Evasão a nível de área de estudo – Importância de atributos para o classificador *Naive Bayes* treinado com *downsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	97,60
nota_cn	80,23
nota_mt	74,06
nota_lc	69,89
nota_ch	63,42
ano_concluiu	61,95
st_conclusao	61,53
concluiu_ens_med	59,76
nu_nota_redacao	55,09
mun_esc_dif_prova	54,12
sit_func_esc	52,07
id_localizacao_esc	51,97
mun_res_dif_esc	51,48
id_dependencia_adm_esc	51,29
idade	48,80
uf_esc	43,58
nu_nota_comp5	38,96
nu_nota_comp3	35,10
nu_nota_comp2	33,80

ingressantes do primeiro semestre continuam tendo uma tendência maior a evadir de sua área de estudo. Os alunos que estão em mais de uma IES também. Em relação às notas, o comportamento também continua o mesmo, indicando que alunos com notas acima da média possuem uma tendência maior de evadir quando comparados com alunos que tiram notas abaixo da média. Em relação ao ano e situação de conclusão do ensino médio, ambos apontam que alunos que já concluíram o ensino médio no momento em que fazem o ENEM têm maior tendência a evadir do que alunos que estão se formando no mesmo ano do exame ou em anos posteriores.

Modelos de redes neurais

Observando a Tabela 3.37 e a Figura 3.30, percebe-se que assim como na avaliação a nível de curso, o modelo treinado com *downsampling* possui desempenho superior aos demais modelos. A diferença de desempenho, considerando a sensibilidade, é significativa a 95% de confiança. A configuração selecionada durante o treinamento para esse balanceamento foi a que utiliza 5 nós na camada escondida e decaimento igual a 0.

Considerando a Tabela 3.39, percebe-se que, mais uma vez, que o modelo de rede neural não considerou os mesmos atributos como importantes quando comparado com o

Tabela 3.37: Evasão a nível de área de estudo – Desempenho dos modelos de redes neurais na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	91,67%	17,52%	96,34%
Downsampling	**52,16%**	**62,42%**	**51,52%**
Upsampling	84,11%	39,81%	86,90%

Tabela 3.38: Evasão a nível de área de estudo – Matriz de confusão do modelo de Redes Neurais com *downsampling*.

——	Referência	
Predição	Sim	Não
Sim	196	2413
Não	118	2564

Tabela 3.39: Evasão a nível de área de estudo – Importância de atributos para o classificador de rede neural treinado com *downsampling*.

Nome do atributo	Importância
doc_integ_sem_de	100,00
ano_concluiu	98,66
ano_ing	94,01
nota_mt	92,55
nu_nota_redacao	82,70
nota_lc	72,73
q037B	72,52
uf_prova23	70,41
candidatos	70,00
q035B	69,00
in_guia_interpretebranco	67,02
nu_nota_comp4	65,34
q003P	65,14
q032I	64,82
q030H	64,31
doc_temp_parcial	64,28
ano_ces2014	63,25
nu_nota_comp1	61,89
nota_cn	61,79
co_ocde_area_geral8	60,33

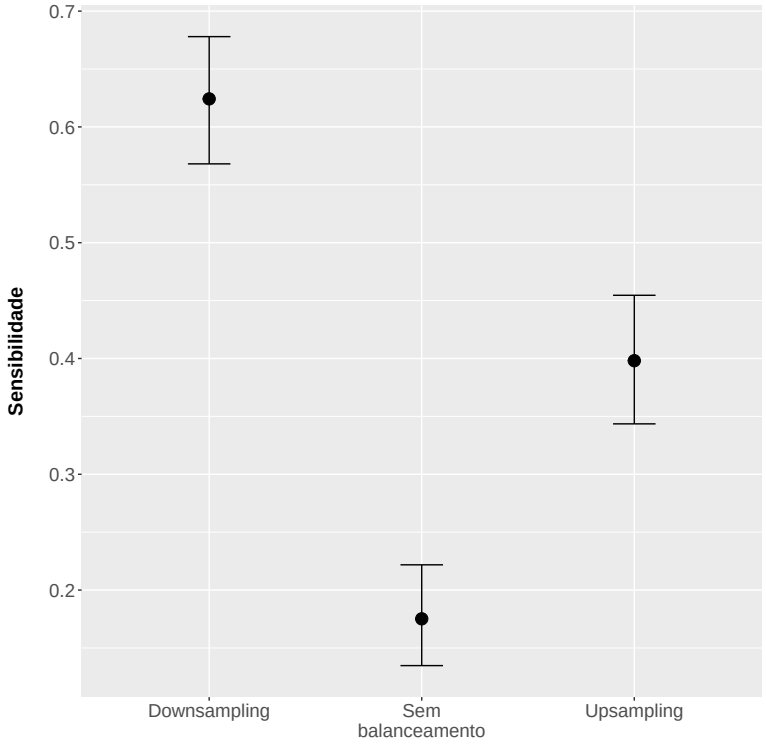

Figura 3.30: Evasão a nível de área de estudo – Intervalos de confiança para a sensibilidade medida na *tabela* de teste para os modelos de redes neurais.

Naive Bayes. Na análise da evasão a nível de curso, o modelo de rede neural foi o único que destoou em relação à importância dos atributos. Para evasão a nível de área de estudo, o modelo selecionado considerou a quantidade de docentes sem dedicação exclusiva, o ano de conclusão do ensino médio, o ano de ingresso no ensino superior, a nota na prova objetiva de matemática e a nota na redação como os atributos mais importantes para distinguir uma evasão de uma retenção em uma área de estudo.

Modelos de regressão logística

Conforme Tabela 3.40, percebe-se que o modelo treinado com *downsampling* obteve o melhor resultado em sensibilidade. No entanto, a diferença para o modelo com *upsampling* não foi grande e, com o auxílio da Figura 3.31, nota-se que ela não foi significativa a 95% de confiança. Sendo assim, fez-se necessário comparar as outras métricas também. O modelo treinado com *upsampling* possui um desempenho melhor nas outras métricas e a diferença foi significativa tanto para acurácia quanto para especificidade. Portanto, o modelo que obteve o melhor desempenho foi o treinado com *upsampling*.

Observando a Tabela 3.42, verifica-se que os dois atributos considerados mais importantes são os mesmos do modelo de *Naive Bayes*. A interpretação deles é a mesma, o aluno

Tabela 3.40: Evasão a nível de área de estudo – Desempenho dos modelos de regressão logística na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	94,18%	10,19%	99,48%
Downsampling	69,00%	66,24%	69,18%
Upsampling	**73,99%**	**65,92%**	**74,50%**

Tabela 3.41: Evasão a nível de área de estudo – Matriz de confusão do modelo de Regressão Logística com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	207	1269
Não	107	3708

Tabela 3.42: Evasão a nível de área de estudo – Importância de atributos para o classificador de regressão logística treinado com *upsampling*.

Nome do atributo	Importância
semestre_ingresso2	100,00
num_ies	79,97
ano_ing	56,51
co_ocde_area_geral5	30,8
nota_cn	26,86
in_compl_monitoria1	23,15
idade	20,26
nota_mt	19,95
candidatos	19,42
ano_ces2011	19,06
nota_ch	18,57
in_compl_estagiobranco	18,41
in_certificadobranco	17,54
q0275	17,17
q0285	15,81
nu_nota_redacao	15,26
q041B	14,95
q0463	14,85
doc_brasileiro	14,73
tp_escolabranco	14,71

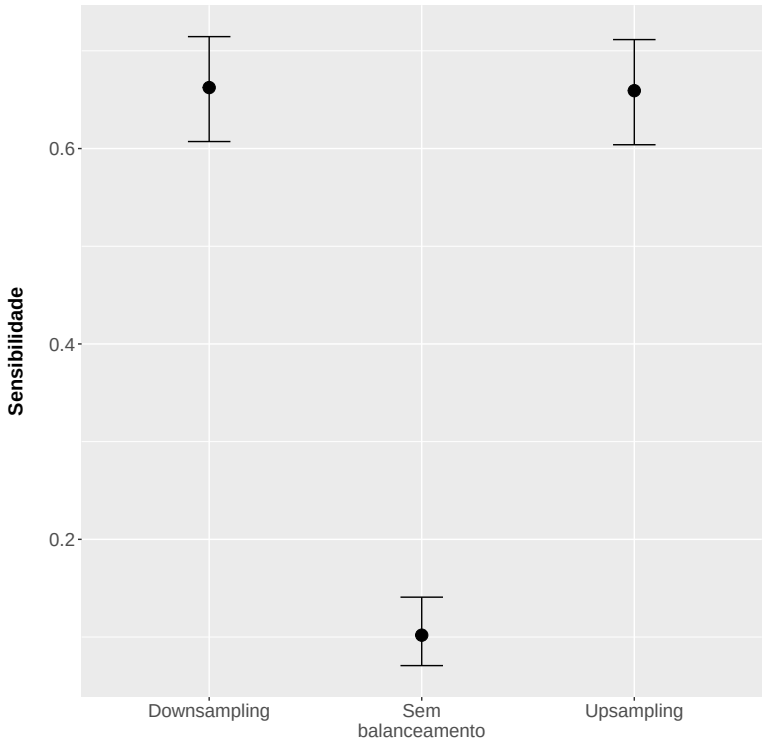

Figura 3.31: Evasão a nível de área de estudo – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de regressão logística.

que ingressa no segundo semestre tem uma chance 5 vezes menor de evadir se comparado a um aluno que ingressa no primeiro semestre. O modelo também indica que existe uma chance 3 vezes maior de evadir para o aumento de cada unidade no atributo de quantidade de IES distintas a que o aluno tenha vínculo. Em relação ao ano de ingresso no ensino superior, o modelo indica que para cada unidade acrescida ao atributo a chance de evadir é 3 vezes menor, ou seja, alunos que ingressaram em 2014 ou 2013 evadem menos de suas áreas de estudo. O modelo também captou uma das áreas como atributo que ajuda a classificar alunos evadidos: área 5 (Engenharia, Produção e Construção). O modelo indica que alunos pertencentes a essa área possuem 3 vezes menos chance de evadir se comparados a alunos da área 1 (Educação). Em todas as interpretações de razões de chances, supõe-se que os outros atributos se mantenham constantes entre os casos comparados.

Modelos CART

O CART gerou o melhor modelo na avaliação da evasão a nível de curso. Porém, analisando a Tabela 3.43, verifica-se que o modelo CART para evasão a nível de área de estudo não obteve o mesmo desempenho para sensibilidade. O modelo treinado com *downsampling*, apesar de parecer ter o melhor desempenho, não possui diferença estatis-

Tabela 3.43: Evasão a nível de área de estudo – Desempenho dos modelos CART na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	92,23%	9,24%	97,47%
Downsampling	49,82%	77,39%	48,08%
Upsampling	**73,28%**	**71,34%**	**73,40%**

Tabela 3.44: Evasão a nível de área de estudo – Matriz de confusão do modelo de CART com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	224	1324
Não	90	3653

ticamente significativa a 95% de confiança em relação ao modelo treinado, conforme pode se observar na Figura 3.32. Comparando as demais métricas, o modelo com *upsampling* mostrou-se superior ao modelo com *downsampling* e, portanto, é o melhor modelo CART para evasão a nível de área de estudo. A configuração desse modelo possui o parâmetro de complexidade igual a 0,0046.

O modelo CART deu importância a diferentes atributos dos vistos nos algoritmos passados. O semestre de ingresso apareceu apenas como o quarto atributo mais importante, enquanto outros dois atributos apareceram pela primeira vez entre os cinco mais importantes: Unidade da Federação em que o aluno estuda no ensino médio ("uf_esc") e a situação de conclusão do ensino médio. Apesar de o atributo "uf_esc" aparecer com relevância, ele não faz parte da árvore gerada pelo modelo, descrita na Figura 3.33.

Na árvore, todos os galhos que levam a esquerda são positivos em relação à divisão realizada. Por exemplo, a primeira divisão é realizada no semestre de ingresso igual a 2. 'N' indica casos de retenção, enquanto 'S' indica casos de evasão. Os números abaixo das letras indicam a probabilidade do caso ser evasão ou retenção e o último número indica o percentual de observações que foi para aquele determinado nó.

Percebe-se que os alunos que ingressam no segundo semestre têm baixa probabilidade de evasão e que eles só são classificados como evasão nos casos em que o número de candidatos para ingressar no curso são menores que 4.268. Para alunos que ingressam no primeiro semestre, a análise é distinta. Após dividir os alunos pelo número de IES distintas e classificar os que possuíam mais de 1,5 como evasão, dividiu-se as observações pela situação de conclusão do ensino médio. Alunos que se formam no ano que prestam o ENEM ou posteriormente foram marcados como retenção em quase todos os casos. Outra tendência que se percebeu em vários dos modelos analisados foi a questão da nota

Tabela 3.45: Evasão a nível de área de estudo – Importância de atributos para o classificador CART treinado com *upsampling*.

Nome do atributo	Importância
num_ies	100,00
uf_esc	85,51
st_conclusao	84,33
semestre_ingresso	76,79
concluiu_ens_med	57,81
ano_concluiu	57,81
nota_ch	48,57
nota_mt	29,89
sit_func_esc	28,52
nota_lc	26,13
nota_cn	22,61
uf_nasc	16,35
candidatos	13,62
ano_ces	13,03
nu_nota_redacao	9,16
co_ocde_area_geral	8,31
in_compl_monitoria	8,07
in_compl_estagio	8,04
ano_ing	7,47
uf_res	6,62

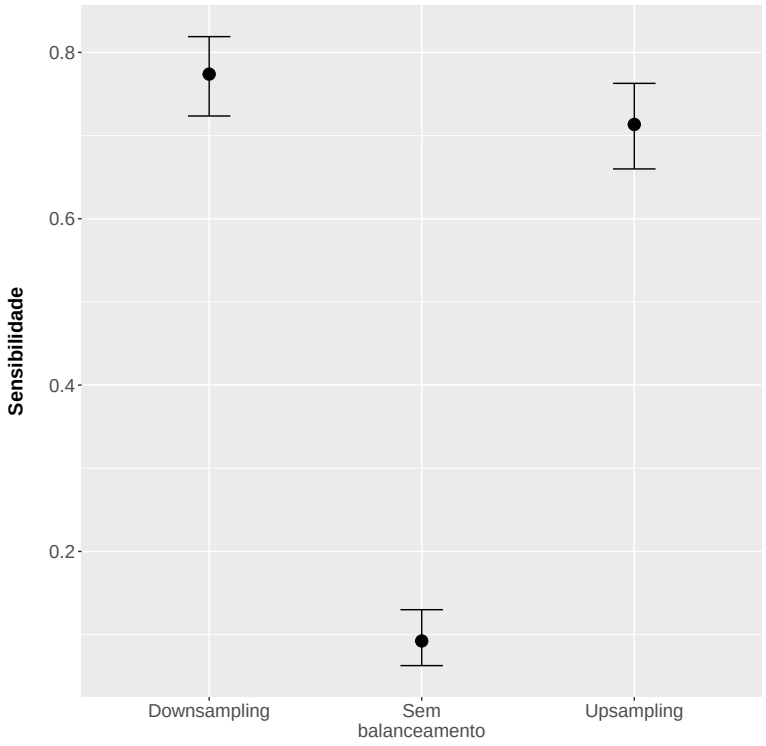

Figura 3.32: Evasão a nível de área de estudo – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos CART.

nas provas objetivas do ENEM. Os alunos que obtiveram notas altas foram classificados com uma probabilidade maior de evasão. Em relação a área de estudo, as áreas 3, 5 e 7 (Ciências Sociais, Negócios e Direito; Engenharia, Produção e Construção e Saúde e Bem estar social) foram classificadas com a menor propensão à evasão.

Modelos C5.0

Na avaliação dos modelos C5.0 para evasão a nível da área de estudo, Tabela 3.46, percebe-se que o comportamento foi similar ao que ocorreu na análise a nível de curso. O modelo treinado com *downsampling*, aparentemente, foio único que não sofreu *overfitting*. O modelo sem balanceamento classificou todas as observações como retenção e o modelo com *upsampling* acertou menos de 20% das observações rotuladas como evasão na *tabela* de teste. Observando a Figura 3.34, infere-se que a diferença do modelo com *downsampling* foi significativa em relação ao modelo com *upsampling*.

Além dos atributos já comentados nos modelos anteriores, que não possuem diferenças de interpretação, o modelo C5.0 traz na Tabela 3.48 alguns atributos não vistos, como, por exemplo, quatro questões do questionário socioeconômico: q034, q030, q045 e q042.

Tabela 3.46: Evasão a nível de área de estudo – Desempenho dos modelos C5.0 na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	94,07%	0,00%	100,00%
Downsampling	**71,91%**	**70,06%**	**72,03%**
Upsampling	88,26%	18,47%	92,67%

Tabela 3.47: Evasão a nível de área de estudo – Matriz de confusão do modelo de C5.0 com *downsampling*.

——	Referência	
Predição	Sim	Não
Sim	220	1392
Não	94	3585

Tabela 3.48: Evasão a nível de área de estudo – Importância de atributos para o classificador C5.0 treinado com *upsampling*.

Nome do atributo	Importância
in_compl_estagio	100,00
nu_integralizacao_integral	100,00
q034	100,00
num_ies	100,00
nota_cn	100,00
nu_nota_redacao	100,00
semestre_ingresso	100,00
in_compl_monitoria	100,00
nota_mt	100,00
uf_esc	100,00
uf_prova	100,00
uf_nasc	99,73
q030	99,42
ano_ing	97,40
candidatos	97,03
q045	96,82
doc_temp_parcial	96,60
nu_integralizacao_noturno	94,06
nu_nota_comp2	93,47
q042	92,89

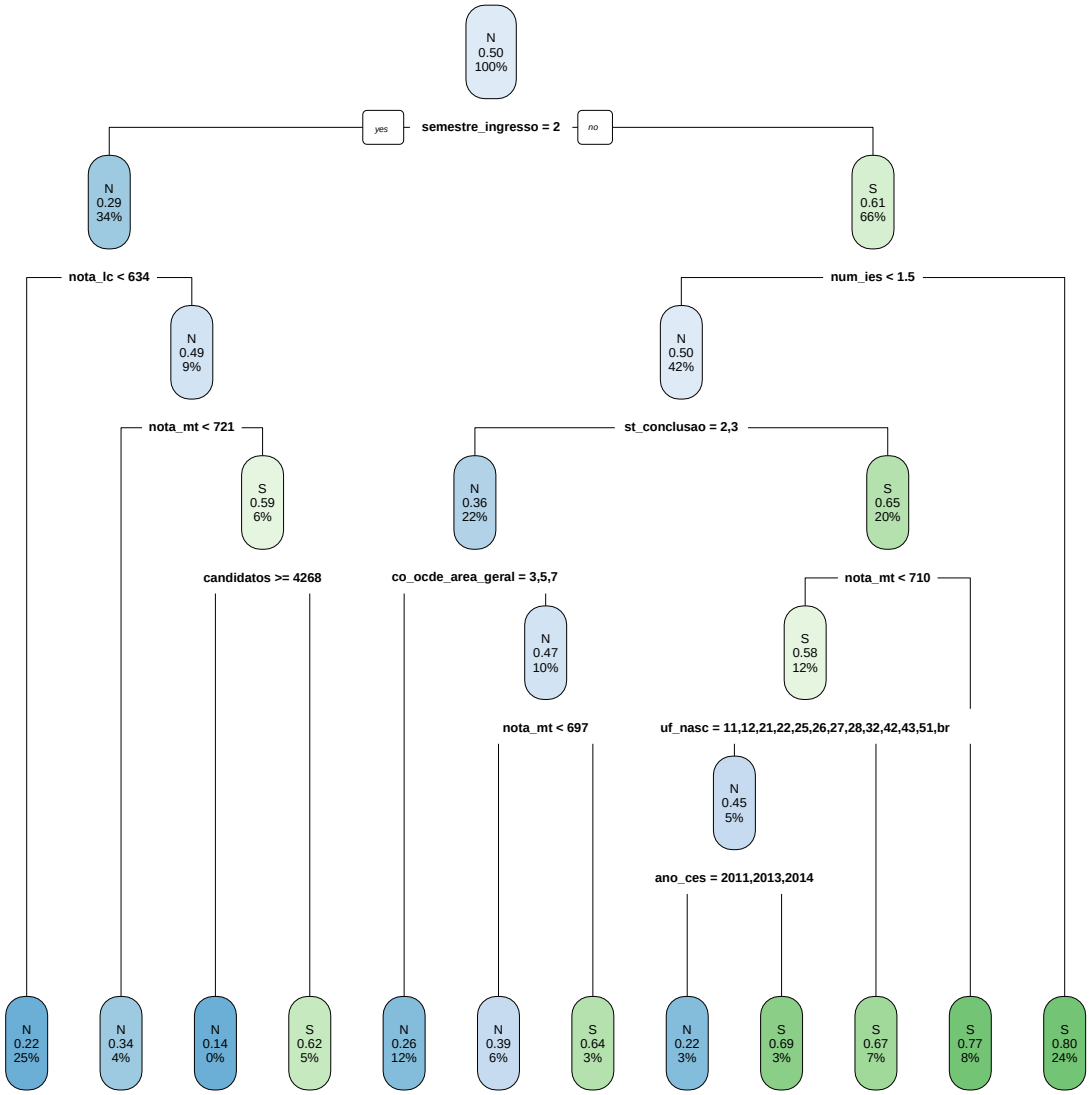

Figura 3.33: Evasão a nível de área de estudo – Árvore gerada pelo modelo CART treinado com *upsampling*.

O modelo indicou que alunos que deixaram de cursar o ensino médio por mais de um ano (q034) possuem maior tendência de evadir de suas áreas de estudo. Já para a quantidade de tempo gasto para se formar no ensino médio (q030), não há uma interpretação clara do comportamento do modelo, pois ele misturou alunos que levaram o tempo esperado para se formar, 8 anos, com alunos que levaram três ou mais anos que o esperado na mesma divisão. O mesmo ocorreu com as questões que pediam ao aluno para indicar a importância dele ter começado a trabalhar para adquirir experiência (q045) ou ajudar

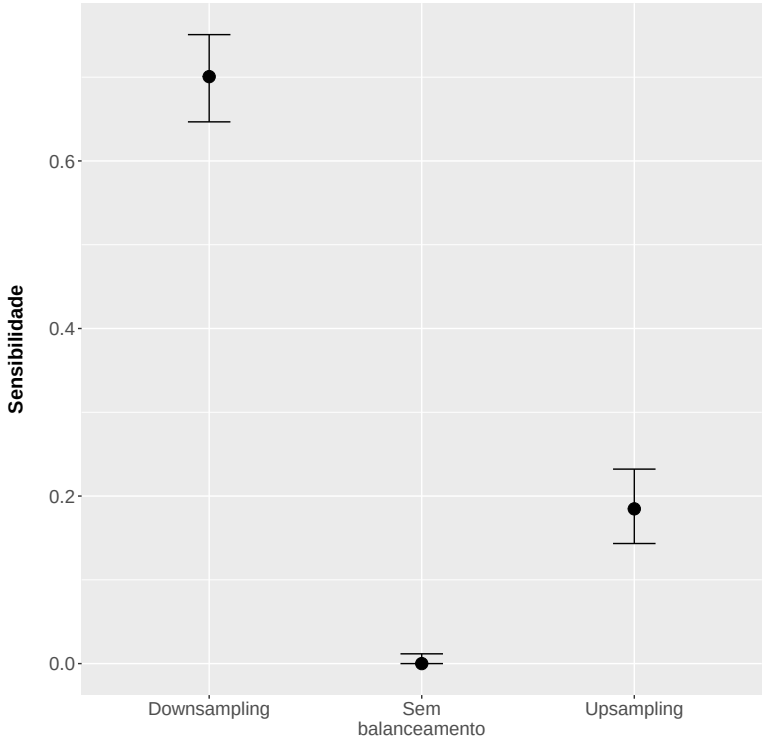

Figura 3.34: Evasão a nível de área de estudo – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos C5.0.

os pais com as despesas de casa (q042). Os demais atributos já interpretados no modelo C5.0 para evasão a nível de curso mantiveram o mesmo comportamento neste modelo.

Comparação entre os algoritmos

Após analisar todos os modelos para evasão a nível de área de estudo, foi feita uma comparação entre os algoritmos utilizados. Mais uma vez, foram escolhidos os modelos com os melhores desempenhos, baseado em sensibilidade dos diferentes algoritmos. Para CART e regressão logística foram utilizados os modelos treinados com *upsampling*. Para os demais algoritmos foram utilizados os treinados com *downsampling*.

A Figura 3.35 mostra a comparação entre os melhores modelos dos diferentes algoritmos através de intervalos de confiança com correção de Bonferroni. Pode-se notar que, para este nível de evasão, nenhum modelo se sobressaiu. Todos tiveram desempenho de sensibilidade similares, sem diferença estatisticamente significativa a 95% de confiança entre nenhum.

Utilizou-se a mesma ideia para determinar o melhor modelo quando a comparação é entre os tipos de balanceamento. As outras métricas foram comparadas e constatou-se que não há um único modelo que se destaque. Há um empate entre CART, C5.0 e regressão

Tabela 3.49: Evasão a nível de área de estudo – Tempo de execução de modelos empatados em desempenho.

Modelo	Tempo de execução
CART	25 segundos
C5.0	40 segundos
Regressão Logística	345 segundos

logística. Sendo assim, utilizou-se o último critério de desempate, o custo computacional. Utilizando a mesma máquina, com processador Intel i5-3570k a 4.2Ghz, 16Gb de memória ram a 1600mhz e utilizando o R 64-bits para Windows 10, foram computados os tempos de execução de acordo com a Tabela 3.49.

Sendo assim, o CART foi o algoritmo que consumiu menos tempo para treinar seus modelos de evasão a nível de área de estudo, mesmo treinando os modelos na *tabela* de *upsampling* contra o C5.0 treinando na *tabela* de *downsampling* e regressão logística também na tabela de *upsampling*.

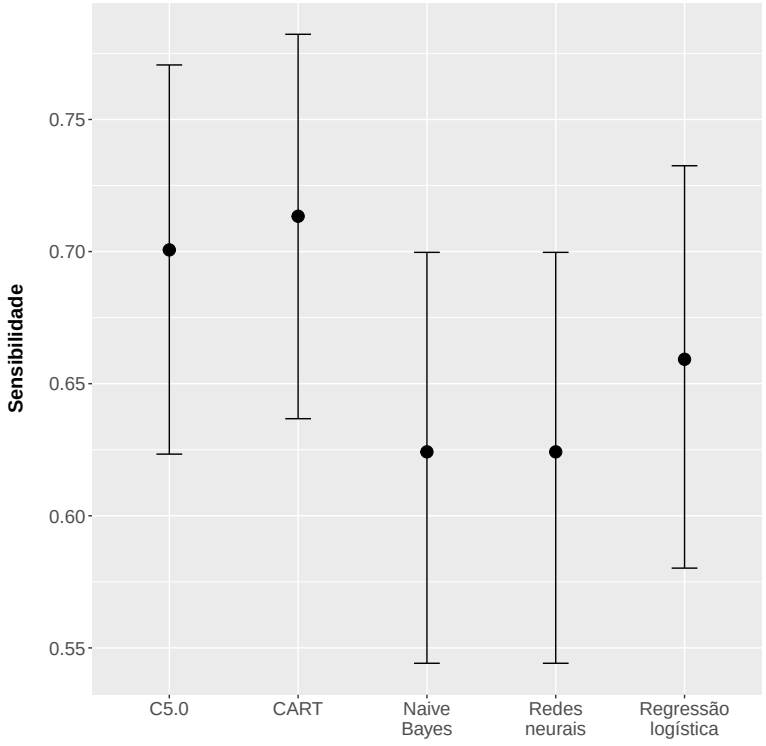

Figura 3.35: Evasão a nível de área de curso – Intervalos de confiança com correção de Bonferroni a 95% para a sensibilidade medida na *tabela* de teste para os melhores modelos de cada algoritmo.

Tabela 3.50: Evasão a nível de IES – Desempenho dos modelos de *Naive Bayes* na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	79,18%	40,73%	81,25%
Downsampling	**66,04%**	**64,57%**	**66,12%**
Upsampling	67,81%	57,62%	68,36%

Tabela 3.51: Evasão a nível de IES – Matriz de confusão do modelo de Naive Bayes com *downsampling*.

——	Referência	
Predição	Sim	Não
Sim	195	1904
Não	107	3716

3.5.3 Evasão a nível de IES

O último nível analisado foi a evasão por IES. Assim como na análise dos modelos de evasão a nível de área de estudo, as tabelas e gráficos com os desempenhos dos modelos durante a fase de treinamento foram apresentados no Apêndice A.

Modelos de *Naive Bayes*

Na Tabela 3.50, nota-se que, assim como em todos os modelos vistos até aqui, o modelo treinado sem balanceamento teve desempenho inferior aos modelos balanceados. Dentre os modelos balanceados, o com *downsampling*, conforme resultados, obteve um desempenho superior. No entanto, ao verificar a Figura 3.36, percebe-se que não há uma diferença estatisticamente significativa entre a sensibilidade do modelo com *downsampling* para a do modelo com *upsampling*. Dessa forma, para determinar o melhor modelo, foram verificadas as demais métricas. Na Tabela 3.50, verifica-se que o modelo com *downsampling* obteve o melhor resultado, no entanto, ao analisar os intervalos de confiança, nota-se que houve empate em todas as métricas apresentadas. Portanto, o melhor modelo de *Naive Bayes* para evasão a nível de IES foi o treinado com *downsampling*, pois foi o de menor custo computacional.

Analisando a Tabela 3.52, percebe-se que os atributos são praticamente os mesmos vistos nas análises de evasão a nível de curso e de área de estudo. Sua interpretação também foi muito similar. Não há novidades a respeito dos modelos de *Naive Bayes* para evasão a nível de IES em relação às análises anteriores.

Tabela 3.52: Evasão a nível de IES – Importância de atributos para o classificador de *Naive Bayes* treinado com *downsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	93,46
nota_ch	90,75
nota_cn	82,59
nota_mt	78,70
nota_lc	77,26
st_conclusao	65,47
ano_concluiu	65,27
concluiu_ens_med	62,25
nu_nota_redacao	60,11
idade	56,78
sit_func_esc	50,45
id_localizacao_esc	50,19
id_dependencia_adm_esc	48,56
mun_esc_dif_prova	48,19
mun_res_dif_esc	47,39
uf_esc	41,89
uf_nasc	39,61
ano_enem	37,75
tp_sexo	35,34

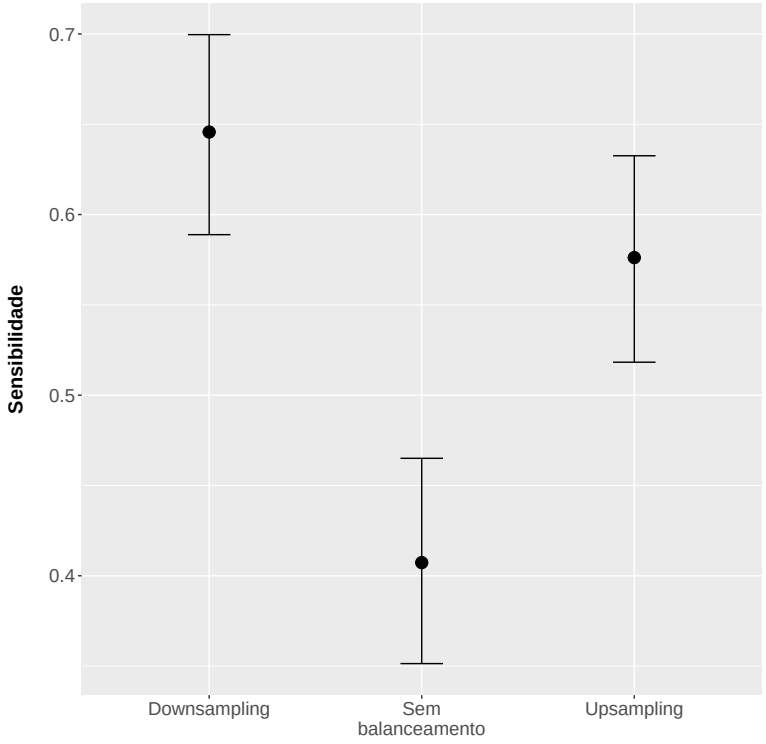

Figura 3.36: Evasão a nível de IES – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos de *Naive Bayes*.

Modelos de redes neurais

Diferentemente do que ocorreu ao analisar os modelos de redes neurais para evasão a nível de curso e área de estudo, o modelo com *upsampling* e *downsampling* tiveram desempenho similar para sensibilidade, como pode ser visto na Tabela 3.53. Como o desempenho foi numericamente idêntico, foi trivial perceber que não houve diferença significativa entre os modelos, e portanto, não há a necessidade de apresentar a Figura com os intervalos de confiança. Ao utilizar as demais métricas para desempatar, o modelo com *upsampling* mostrou desempenhos superiores ao de *downsampling* com diferenças estatisticamente significantes a 95% de confiança.

Tabela 3.53: Evasão a nível de IES – Desempenho dos modelos de redes neurais na *tabela* de teste por tipo de balanceamento.

———	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,48%	11,59%	97,88%
Downsampling	59,63%	55,96%	59,82%
Upsampling	**72,04%**	**55,96%**	**72,90%**

Tabela 3.54: Evasão a nível de IES – Matriz de confusão do modelo de Redes Neurais com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	169	4097
Não	133	1523

Tabela 3.55: Evasão a nível de IES – Importância de atributos para o classificador de rede neural treinado com *upsampling*.

Nome do atributo	Importância
doc_graduacao	100,00
doc_temp_parcial	76,14
doc_exercicio	62,07
ano_ces2014	55,11
doc_brasileiro_nat	54,84
ano_ces2013	50,22
doc_mestrado	45,46
ano_ces2012	45,41
semestre_ingresso2	45,04
idade	43,45
q030B	41,00
q002G	40,87
doc_integ_de	40,84
co_ocde_area_geral2	39,57
vagas	37,76
q035C	36,21
q0234	35,76
num_ies	34,44
doc_especializacao	31,73
q0254	31,58

Tabela 3.56: Evasão a nível de IES – Desempenho dos modelos de regressão logística na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	94,60%	7,62%	99,27%
Downsampling	68,47%	66,89%	68,56%
Upsampling	**76,56%**	**67,22%**	**77,06%**

Tabela 3.57: Evasão a nível de IES – Matriz de confusão do modelo de Regressão Logística com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	203	1289
Não	99	4331

Analisando a Tabela 3.55, percebe-se que novamente o modelo de rede neural não considerou como importantes os mesmos atributos que os modelos de outros algoritmos. O semestre de ingresso, por exemplo, está no topo da lista de quase todos os modelos dos outros algoritmos, mas apareceu apenas na nona colocação em termos de importância neste modelo. Assim como o modelo de evasão a nível de área de estudo, o modelo de rede neural deu uma importância maior para características dos professores. O primeiro atributo foi a quantidade de docentes que possuem apenas graduação, o segundo foi a quantidade de docentes que trabalham em tempo parcial e o terceiro contabilizou a quantidade de docentes que estavam em exercício, ou seja, que não estavam afastados para capacitação ou por algum problema de saúde. O modelo também considerou a quantidade de docentes naturalizados e a quantidade de docentes que possuíam mestrado como escolaridade máxima.

Modelos de regressão logística

Na Tabela 3.56, nota-se que os desempenhos de sensibilidade dos modelos de *downsampling* e *upsampling* foram similares. Assim como nos modelos de redes neurais, não foi necessário verificar os intervalos de confiança da sensibilidade para perceber que não houve diferença estatisticamente significativa entre os modelos com *downsampling* e *upsampling*. Os resultados obtidos nessa tabela foram praticamente iguais aos obtidos na avaliação de evasão a nível de área de estudo para regressão logística. Mais uma vez, no desempate, o modelo com *upsampling* registrou desempenho superior ao modelo com *downsampling*. A diferença nas demais métricas foi estatisticamente significativo a 95% de confiança.

Ao observar a Tabela 3.58, percebe-se que de forma similar à avaliação do modelo de regressão logística para evasão a nível de área de estudo, os três atributos mais importan-

Tabela 3.58: Evasão a nível de IES – Importância de atributos para o classificador de regressão logística treinado com *upsampling*.

Nome do atributo	Importância
num_ies	100,00
semestre_ingresso2	90,45
ano_ing	65,46
q041B	24,77
in_compl_monitoria1	24,22
nota_ch	23,98
nota_cn	23,13
ano_ces2012	21,89
q0275	20,69
ano_enem2011	20,28
q0253	19,40
co_ocde_area_geral2	18,62
ano_ces2011	18,03
co_ocde_area_geral1	17,49
candidatos	17,43
q0255	17,31
q0047	17,14
co_ocde_area_geral8	17,12
co_ocde_area_geral4	16,14
q020B	16,02

Tabela 3.59: Evasão a nível de IES – Desempenho dos modelos CART na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,80%	8,94%	98,36%
Downsampling	66,43%	72,52%	66,10%
Upsampling	**75,68%**	**68,21%**	**76,09%**

Tabela 3.60: Evasão a nível de IES – Matriz de confusão do modelo de CART com *upsampling*.

——	Referência	
Predição	Sim	Não
Sim	206	1344
Não	96	4276

tes não mudaram. No entanto, a ordem entre eles foi modificada. Os outros atributos, da quarta posição em diante, continuaram obtendo aproximadamente metade da importância dada ao terceiro atributo mais importante. A interpretação dos atributos mais importantes não foi diferente da explicada nos modelos de regressão logística analisados nos outros níveis de evasão.

Modelos CART

A Tabela 3.59 mostra que, mais uma vez, os modelos treinados com balanceamento obtiveram desempenho superior em sensibilidade se comparados com o modelo treinado sem balanceamento. O modelo com *downsampling* exibiu um desempenho 4% melhor que o de *upsampling* na métrica de interesse. No entanto, ao analisar a Figura 3.37, conclui-se que essa diferença não é estatisticamente significativa a 95% de confiança. Dessa forma, foi novamente necessário desempatar os modelos balanceados em termos de desempenho. Nas métricas de acurácia e especificidade, o modelo com *upsampling* foi superior ao modelo com *downsampling* e a diferença em ambas as métricas foi estatisticamente significativa a 95%.

Comparando-se a Tabela 3.61 com a Tabela 3.45, percebe-se que os atributos considerados importantes mudaram. Enquanto na análise da evasão a nível de área de estudo havia quatro atributos com valores de importância relativa superior a 75, na evasão a nível de IES apenas um atributo registrou tal valor, o semestre de ingresso.

A Figura 3.38 mostra a árvore formada pelo modelo CART treinado com *upsampling*. A primeira divisão foi realizada com o número de IES distintas a que um aluno está vinculado. No entanto, esse atributo não foi considerado tão importante quanto o semestre de ingresso dos alunos. Novamente, a interpretação foi a mesma dos modelos anteriores:

Tabela 3.61: Evasão a nível de IES – Importância de atributos para o classificador CART treinado com *upsampling*.

Nome do atributo	Importância
semestre_ingresso	100,00
num_ies	59,30
nota_ch	49,86
uf_esc	49,36
nota_lc	44,75
nota_cn	41,61
nota_mt	34,63
ano_concluiu	31,05
st_conclusao	29,96
concluiu_ens_med	29,56
id_dependencia_adm_esc	27,94
uf_res	14,98
q003	14,17
uf_prova	12,94
ano_ing	11,94
ano_ces	11,94
candidatos	11,86
in_compl_pesquisa	8,14
in_compl_estagio	8,14
in_compl_monitoria	8,14

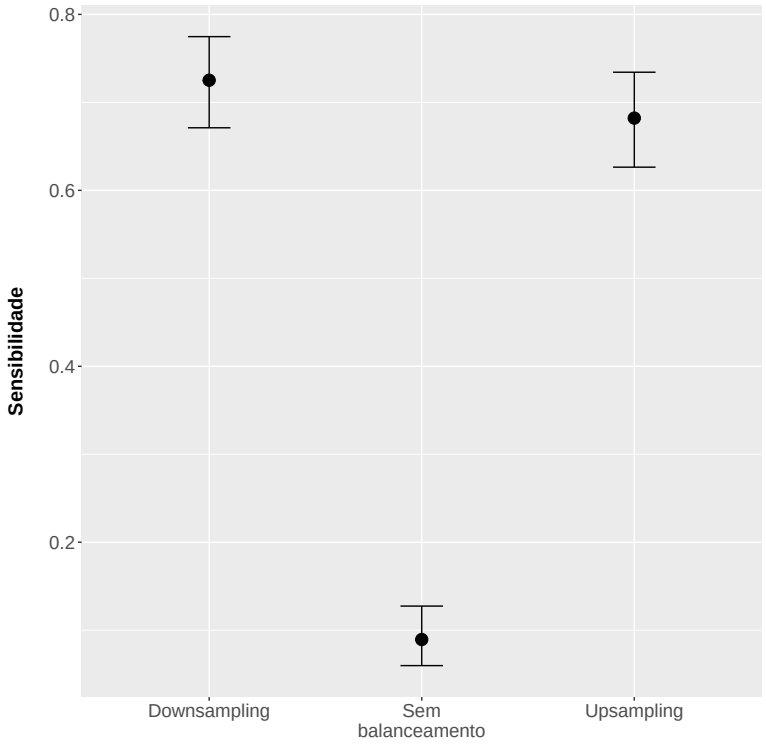

Figura 3.37: Evasão a nível de IES – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos CART.

alunos ingressantes no primeiro semestre têm uma propensão maior de evadir quando comparados a alunos ingressantes no segundo semestre. Além disso, alunos com notas acima da média nas provas objetivas apresentaram maior tendência a evadir em relação aos com notas abaixo da média. Alunos que ainda não eram formados no momento em que fizeram o ENEM têm menor propensão a evadir quando comparados a alunos que já se formaram.

Modelos C5.0

Conforme Tabela 3.62, existem semelhanças em relação aos desempenho dos modelos C5.0 nas análises da evasão a nível de área de estudo e curso. Na classificação da *tabela* de teste, mais uma vez, o modelo com *downsampling* se destacou. Seu desempenho para sensibilidade foi superior ao dos demais modelos, mesmo quando considerado o intervalo de confiança, visto na Figura 3.39.

Analisando a Tabela 3.64, nota-se que muitos dos atributos vistos anteriormente em modelos C5.0 se repetem. Das questões do questionário socioeconômico que apareceram nos modelos C5.0 anteriores, as q031 e q041 são novas e só aparecem entre os 20 atributos mais importantes para a evasão a nível de IES. A q031 pergunta ao aluno se ele deixou

Tabela 3.62: Evasão a nível de IES – Desempenho dos modelos C5.0 na *tabela* de teste por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	94,90%	0,00%	100,00%
Downsampling	**73,81%**	**71,52%**	**73,93%**
Upsampling	89,55%	23,51%	93,10%

Tabela 3.63: Evasão a nível de IES – Matriz de confusão do modelo de C5.0 com *downsampling*.

——	Referência	
Predição	Sim	Não
Sim	216	1465
Não	86	4155

Tabela 3.64: Evasão a nível de IES – Importância de atributos para o classificador C5.0 treinado com *downsampling*.

Nome do atributo	Importância
in_compl_monitoria	100,00
in_compl_pesquisa	100,00
num_ies	100,00
semestre_ingresso	100,00
uf_esc	100,00
uf_prova	100,00
nota_ch	100,00
q031	100,00
uf_nasc	99,67
ano_ing	99,56
nota_mt	98,95
nota_cn	98,68
q045	98,51
q034	98,46
nota_lc	97,80
in_compl_estagio	97,74
q026	96,97
q041	95,38
idade	94,55
candidatos	93,84

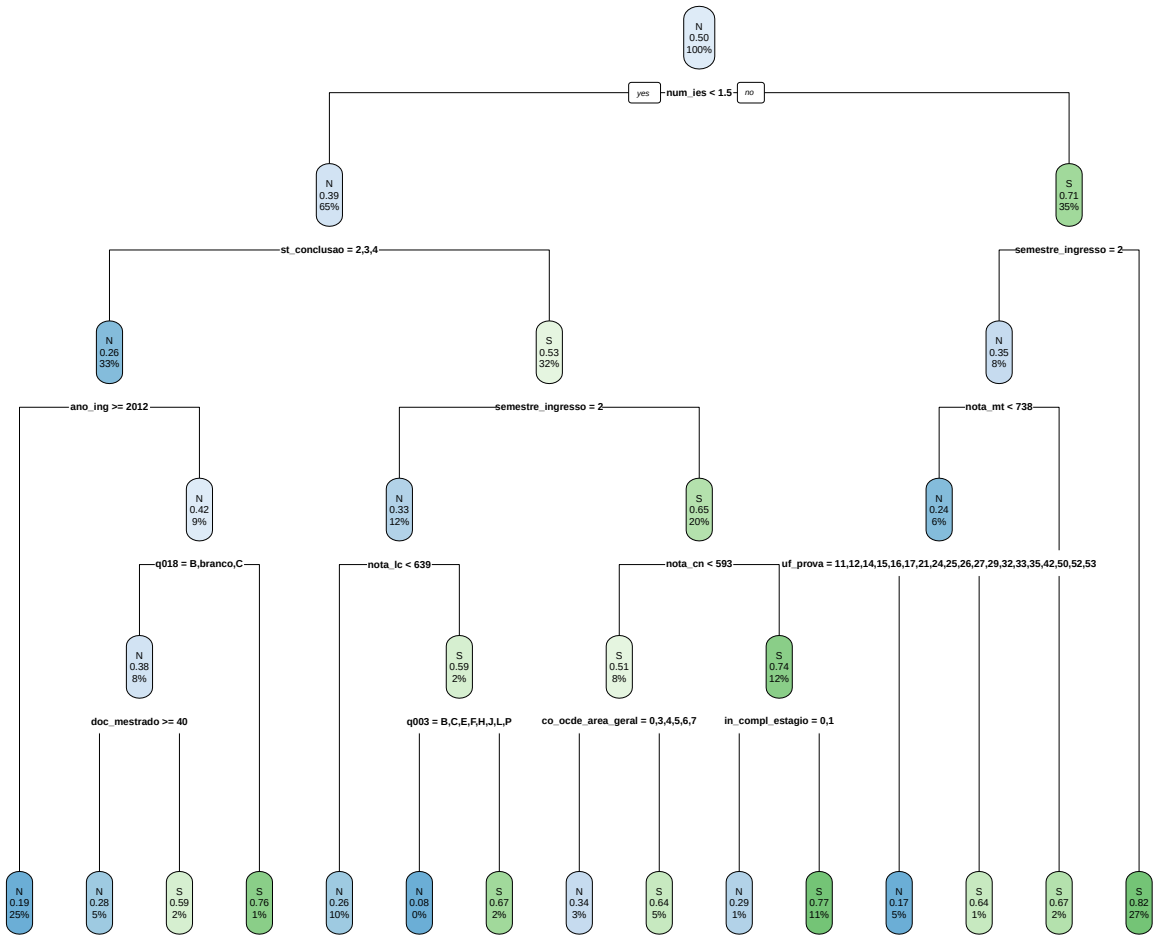

Figura 3.38: Evasão a nível de IES – Árvore gerada pelo modelo CART treinado com *upsampling*.

de estudar durante o ensino fundamental. Enquanto a q041 pergunta o número de horas semanais trabalhadas. Assim como no modelo C5.0 para evasão a nível de área de estudo, as questões do questionário socioeconômico não foram agrupadas de forma a permitir uma interpretação clara da perspectiva de negócio. Exemplificando, para a questão sobre o tempo de trabalho durante a semana, o modelo agrupou em uma das árvores os valores extremos em um nó, ou seja, as alternativas 'A' e 'E' e, em outro nó, foram agrupados outros valores, como 'B' e 'D'. Em outra árvore, o modelo agrupou os valores preenchidos com 'A' (menor valor), 'C' (valor do meio) e 'E' (maior valor). O atributo "candidatos", que conta o número de pessoas que se inscreveram no ano para ingressar em determinado

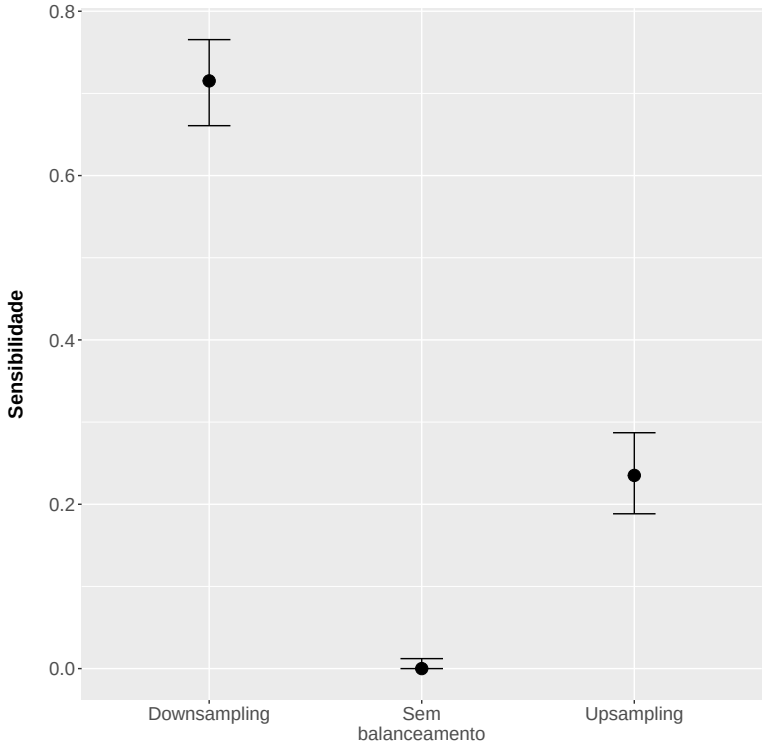

Figura 3.39: Evasão a nível de IES – Intervalos de confiança a 95% para a sensibilidade medida na *tabela* de teste para os modelos C5.0.

curso, possui uma interpretação consistente. Dependendo da posição que ele se encontra em uma árvore, seu valor de divisão muda. No entanto, os alunos que estavam vinculados a cursos que possuíam mais candidatos que o valor da divisão apresentaram tendência a não evadir. Ou seja, em cursos muito concorridos, espera-se que a evasão seja baixa.

Comparação entre os algoritmos

Depois de analisados todos os modelos criados para evasão a nível de IES, fez-se uma comparação entre os desempenhos dos melhores modelos para cada algoritmo. Para os modelos *Naive Bayes* e C5.0 foram os treinados com *downsampling*. Para os demais, os treinados com *upsampling*.

A comparação entre os desempenhos baseada em sensibilidade, dos modelos na Figura 3.40, mostrou um resultado similar ao visto na evasão a nível de área de estudo. Não há um algoritmo que se destaque.

Utilizou-se, então, as outras métricas para buscar um desempate entre o desempenho dos modelos. No entanto, assim como na comparação da evasão a nível de área de estudo, nenhum modelo se sobressaiu. Aconteceu um empate entre CART, C5.0, regressão logística e redes neurais. Dessa forma, utilizou-se o último critério de desempate, o custo

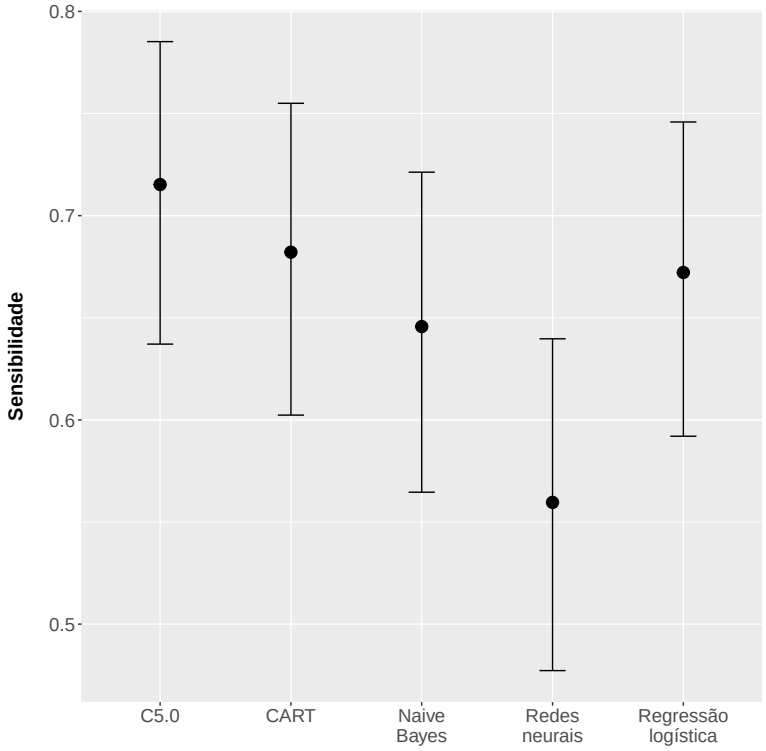

Figura 3.40: Evasão a nível de IES – Intervalos de confiança com correção de Bonferroni a 95% para a sensibilidade medida na *tabela* de teste para os melhores modelos de cada algoritmo.

computacional. A mesma máquina descrita anteriormente foi utilizada, e os resultados de tempo de execução podem ser vistos na Tabela 3.65.

O CART foi, mais uma vez, o algoritmo que consumiu menos tempo para treinar seus modelos. Portanto, para evasão a nível de IES, o modelo CART foi considerado o melhor.

3.5.4 Revisão do processo e determinação da próxima etapa

Nesta etapa foram descritos os modelos dos cinco algoritmos propostos para os três diferentes níveis de evasão definidos. Na interpretação de como os atributos contribuíam para

Tabela 3.65: Evasão a nível de IES – Tempo de execução de modelos empatados em desempenho.

Modelo	Tempo de execução
CART	28 segundos
Regressão Logística	332 segundos
C5.0	1.284 segundos
Redes Neurais	10.230 segundos

Tabela 3.66: Tipo de balanceamento escolhido como o melhor para cada algoritmo e nível de evasão.

Algoritmos	Curso	Área de estudo	IES
Naive Bayes	*upsampling*	*downsampling*	*downsampling*
Redes Neurais	*downsampling*	*downsampling*	*upsampling*
Regressão logística	*upsampling*	*upsampling*	*upsampling*
CART	*downsampling*	*upsampling*	*upsampling*
C5.0	*downsampling*	*downsampling*	*downsampling*

a classificação dos alunos na classe evasão, percebeu-se que os modelos treinados pelos mesmos algoritmos são similares entre si, mesmo para os diferentes níveis. Os modelos de um algoritmo tendem a utilizar os mesmos atributos e as interpretações não são alteradas.

Entre os algoritmos, há diferenças. Os modelos de redes neurais e C5.0 foram os que mais destoaram. Eles utilizaram diferentes atributos para classificar os alunos. No entanto, as interpretações dos atributos em comum foram similares. Em termos de desempenho, não há evidências de que um dos algoritmos seja melhor para este tipo de problema e nem que algum dos níveis definidos para evasão contribua para o aumento de desempenho. Quando analisada a evasão a nível de curso, o modelo CART mostrou um desempenho superior em relação aos outros algoritmos. No entanto, nos níveis de área de estudo e IES isso não ocorreu.

Entre as formas de balanceamento, os resultados mostram que modelos treinados sem balanceamento das classes não são capazes de identificar alunos rotulados como evadidos. A Tabela 3.66 mostra que nenhum modelo treinado sem balanceamento foi escolhido como o melhor modelo para uma determinada nível e algoritmo.

As principais características identificadas nos alunos que possuem maior propensão a evadir foram: ingresso no primeiro semestre, possuir vínculos com mais de uma IES e obter notas acima da média nos exames do ENEM, sendo que não há uma área do conhecimento nos exames que seja proeminente. Além disso, foram vistas outras características que ajudam a classificar um aluno como evasão, por exemplo, ele já ter concluído o ensino médio ao fazer as provas do ENEM.

Na confecção e avaliação inicial dos modelos, foi necessário em diversas ocasiões voltar e revisar a construção das *tabelas* utilizadas, assim como prevê o processo de mineração. O texto apresentado neste capítulo foi resultado desse ciclo de modelagem e preparação dos dados. Seguindo o processo do CRISP-DM, nesta etapa deve ser determinado o próximo passo, uma vez que os modelos já foram analisados. Logo, na seção seguinte será detalhada a etapa de implementação.

3.6 Implementação

Nesta etapa, os dados devem ser organizados e apresentados de forma que o usuário final possa utilizá-los. Dependendo dos requerimentos, este passo pode ser simples como gerar um relatório ou complexo como implementar uma automatização de todo o processo. A implementação é dividida em:

- planejar a implementação;
- planejar o monitoramento e manutenção e
- geração do relatório final.

A implementação do projeto é feita em duas partes. A primeira é a confecção deste texto, que descreve todos os passos e procedimentos realizados do processo de descoberta de conhecimento sobre evasão no ensino superior para qualquer IES do Brasil. Além disso, o texto traz uma introdução ao problema da evasão no ensino superior e um capítulo de revisão de literatura, que aborda trabalhos com enfoques parecidos com os deste trabalho e descrições do funcionamento dos algoritmos utilizados.

A segunda parte é o pacote para o R, *HEdropout*[19], que segue os procedimentos descritos neste texto com a possibilidade de ajustes para que um usuário futuro possa, não apenas repetir os experimentos mostrados neste texto, como também realizar novos experimentos. O pacote permite gerar classificadores para qualquer IES do Brasil, uma vez que o usuário tenha acesso aos dados do Inep ou crie *tabelas* com informações similares. Ele permite escolher quais algoritmos devem ser utilizados para treinar novos modelos, assim como o tipo de balanceamento, o número de validações cruzadas a serem realizadas, o percentual de observações que devem ser alocadas nas *tabelas* de treinamento e teste e o número de unidades de processamento da máquina a serem utilizados para executar o programa.

O pacote é composto por três funções distintas que englobam todo o código utilizado neste trabalho. A primeira, *create_dropout_database*, gera a *tabela* com as informações do CES e ENEM. Ela possui duas opções a serem escolhidas pelo usuário, o tipo de evasão, ou seja, a nível de curso, área de estudo ou IES, e os códigos das IES que devem ser selecionadas para formar a *tabela*. Podem ser escolhidas uma ou mais IES para compor a *tabela*. A segunda, *preprocess*, faz um pré-processamento da *tabela* gerada para realizar o treinamento. Não há opções nesta função. A terceira função, *classify_dropout*, é a principal função do pacote e realiza o treinamento de um novo modelo de acordo com as especificações escolhidas pelo usuário. Esta função também contém o código da função

[19] `https://github.com/lucke71/Classify_dropout/tree/master/HEdropout`

preprocess como uma opção, ou seja, se o usuário for treinar apenas um modelo, ele pode fazê-lo sem utilizar a função de pré-processamento. Caso ele deseje treinar mais de um modelo, o uso da função de pré-processamento é encorajado, uma vez que com ela é possível realizar as tarefas de transformações dos dados apenas uma vez, ao invés de realizá-la todas as vezes que se for treinar um modelo novo no mesmo grupo de dados.

A manutenção e monitoramento do pacote será feita através das ferramentas disponíveis no ambiente do *github*. Nesse ambiente é possível que usuários da comunidade abram reclamações relacionadas ao uso do pacote ou sugiram melhorias. As sugestões e relatórios de problemas no pacote ficam registrados na página do projeto e então é possível corrigi-los ou implementar as melhorias sugeridas.

Capítulo 4

Conclusão e trabalhos futuros

Apesar da divulgação sistemática de dados sobre o ensino superior brasileiro pelo Inep, não há divulgação oficial consolidada de informações ou de indicadores sobre a evasão. Entretanto, sabe-se que é um assunto de crescente interesse ao redor do mundo [1, 2, 3]. No Brasil, apesar de estudos incipientes, acredita-se que a tendência será a mesma, uma vez que a interrupção do ciclo de estudos pode gerar inúmeras consequências, como, por exemplo, a perda de recursos do Estado, o prejuízo financeiro e os prejuízos social e individual.

Mapear as causas deste problema para tentar tratá-las, pode trazer contribuições efetivas a longo prazo, uma vez que é verificado que o nível educacional de um país pode influenciar positivamente sua economia das mais diversas formas [16]. Este trabalho buscou identificar características comuns a alunos que evadem do ensino superior brasileiro, verificar a eficácia de diferentes algoritmos de seleção, analisar casos específicos de grupos de alunos classificados como evasão e propor um *framework* para treinar diferentes classificadores.

O universo para análise escolhido foi o de alunos ingressantes da UnB que participaram das provas do ENEM nos anos 2010 a 2014. Para estes dados, foram adotadas todas as etapas do modelo de processos CRISP-DM. Vale ressaltar, que fez-se necessário voltar e revisar inúmeras vezes a construção das tabelas utilizadas, ciclo previsto no processo de mineração de dados.

Foram definidos, então, três níveis para estudo da evasão: curso, área de estudo e IES. Para cada nível, foram testados cinco algoritmos: *Naive Bayes*, redes neurais, regressão logística, CART e C5.0. E para cada algoritmo, treinados sob três tipos de balanceamento: sem balanceamento, *downsampling* e *upsampling*. Ou seja, no total, este trabalho comparou resultados de 45 modelos.

Analisando o desempenho dos algoritmos escolhidos, o resultado foi que para a evasão a nível de curso, o melhor algoritmo foi o CART, pois ele obteve o melhor desempenho

ao classificar os alunos rotulados como evadidos. A nível de área de estudo e IES não foi identificado um algoritmo que se sobressaísse em relação aos outros. Além disso, nenhum dos modelos escolhidos como o melhor na fase de treinamento foi um modelo treinado sem balanceamento. Mostrando que apesar da classe não possuir um desbalanceamento extremo, o balanceamento das *tabelas* de treinamento se mostrou fundamental para obter modelos que fossem capazes de identificar alunos evadidos. As avaliações dos modelos focou na questão de quão bem eles generalizam. Dessa forma, os desempenhos obtidos ao classificar os dados de teste foram centrais no debate de desempenho dos modelos analisados.

Em relação às características identificadas a partir dos modelos treinados, os alunos que possuem maior propensão a evadir são os que: ingressam no primeiro semestre, possuem vínculos com mais de uma IES e obtêm notas acima da média nos exames do ENEM, sendo que não há uma área do conhecimento nos exames que seja proeminente. Além disso, foram vistas outras características, como o aluno já ter concluído o ensino médio ao fazer as provas do ENEM.

Os resultados obtidos e os que podem ser gerados podem ser utilizados por pesquisadores ou gestores no entendimento do contexto, possibilitando a orientação e/ou implementação de políticas educacionais para a retenção de alunos no ensino superior. Por exemplo, uma determinada IES pode treinar classificadores com os dados disponíveis no Inep e em um ano posterior, utilizar o modelo treinado para identificar entre os alunos ingressantes quais são mais propensos a evadir.

Como proposta de *framework*, o trabalho detalhou todos os procedimentos necessários para a criação da *tabela* que une as informações do CES e ENEM; detalhou os passos necessários para treinar modelos de cinco algoritmos nessa base e resultou no desenvolvimento de um pacote no R para facilitar a execução de trabalhos semelhantes. Diante desta proposta, foi criado um pacote para o R, que possibilita o treinamento de novos modelos, conferindo a um usuário futuro autonomia para selecionar diferentes níveis e critérios de evasão, além da possibilidade de encaminhamento de sugestões de melhoria.

Espera-se que o trabalho realizado gere contribuições efetivas para o debate sobre a evasão, no entanto, sabe-se que ainda há muito a se explorar neste assunto. Ao longo do projeto, foram identificadas possibilidades para trabalhos futuros, como, por exemplo, o treinamento de modelos com mais de duas classes, ou seja, ao invés de rotular alunos apenas como evasão e retenção, rotulá-los com suas situações originais vindas dos dados do CES, como ‘cursando’, ‘desvinculado’, ‘formado’ e assim por diante. Estudar o comportamento de outros algoritmos além dos cinco testados neste trabalho. Realizar ajustes finos nas configurações dos modelos com o intuito de melhorar seus desempenhos. Ainda, focar mais na engenharia de atributos para criar novos atributos capazes de facilitar a

classificação dos modelos treinados. Outra possibilidade é realizar estudos com outros agrupamentos, como alunos de mais de uma IES ou um grupo de IES de um município. Uma outra proposta é a análise de dados de cursos EAD, pois esse tipo de modalidade de ensino tem se popularizado em anos recentes [19] e não faz parte deste trabalho.

Com o acúmulo de dados ao longo dos anos, também pode-se haver uma nova perspectiva sobre este trabalho, possibilitando a análise de determinados grupos de ingressantes e acompanhando suas trajetórias ao longo do tempo.

Referências

[1] Gaioso, Natalícia Pacheco de L.: *A evasão discente na Educação Superior no Brasil: a perspectiva dos dirigentes e dos alunos*. Tese de Mestrado, Universidade Católica de Brasília, 2005. 1, 6, 109

[2] Di Pietro, Giorgio e Cutillo, Andrea: *Degree flexibility and university drop-out: The Italian experience*. Economics of Education Review, 27(5):546–555, outubro 2008, ISSN 0272-7757. http://www.sciencedirect.com/science/article/pii/S0272775707000787, acesso em 2016-06-07. 1, 3, 43, 109

[3] Araque, Francisco e Roldán, Concepción e Salguero Alberto: *Factors influencing university drop out rates*. Computers & Education, 53(3):563–574, novembro 2009, ISSN 0360-1315. http://www.sciencedirect.com/science/article/pii/S0360131509000815, acesso em 2016-06-07. 1, 3, 5, 109

[4] Lobo, Maria Beatriz de Carvalho Melo: *Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções*. Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos, (25), 2012. http://www.institutolobo.org.br/imagens/pdf/artigos/art_087.pdf, acesso em 2016-06-21. 1, 2

[5] Morosini, Marília Costa e Casartelli, Alam de Oliveira e Schimitt Rafael Eduardo e Santos Bettina Steren e Benso Ana Cristina e Gessinger Rosana Maria: *A Evasão na Educação Superior no Brasil: uma análise da produção de conhecimento nos periódicos Qualis entre 2000-2011*. Porto Alegre/RS–Brasil. Faculdade de Educação–FACED. Pontifícia Universidade Católica do Rio Grande do Sul-PUCRS, 10:1–10, 2012. http://clabes-alfaguia.org/clabes-2011/ponencias/ST_1_Abandono/12_MorosiniM_Abandono_ESBrasil.pdf, acesso em 2016-07-19. 1

[6] Tinto, Vincent: *Dropout from higher education: A theoretical synthesis of recent research*. Review of educational research, 45(1):89–125, 1975. http://www.jstor.org/stable/1170024, acesso em 2016-07-06. 1, 2, 5

[7] IBGE: *Histórico do censo da educação superior*. Disponível em http://ces.ibge.gov.br/base-de-dados/metadados/inep/censo-da-educacao-superior.html. 1

[8] Brasil: *Lei nº 9.448. Transforma o Instituto Nacional de Estudos e Pesquisas Educacionais - Inep em Autarquia Federal, e dá outras providências*, março 1997. Disponível em: http://www.planalto.gov.br/Ccivil_03/LEIS/L9448.htm. 1

[9] Brasil: *Decreto nº 6.425. Dispõe sobre o censo anual da educação*, abril 2008. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/Decreto/D6425.htm. 1, 20

[10] Brasil: *Portaria nº 2.255. Aprova o Regimento Interno do Inep*, agosto 2003. Disponível em: https://www.legisweb.com.br/legislacao/?id=184900. 1, 4

[11] Brasil: *Portaria 563. estabelece cronograma e responsabilidades para o preenchimento do censo 2016*, dezembro 2015. Disponível em: http://download.inep.gov.br/educacao_superior/censo_superior/legislacao/2015/portaria_n563_de_17122015_censo_educacao_superior_2015.pdf. 1

[12] Hagedorn, L. S.: *How to define retention: A new look at an old problem. InA. Seidman (Ed.), College student retention: Formula for success (pp. 89-105). Westport, CT: American Council on Higher Education.* Praeger Publishers, 2005. 2, 5

[13] Filho, Silva e Lobo, Roberto Leal e Motejunas Paulo Roberto e Hipolito Oscar e Lobo Maria Beatriz de Carvalho Melo: *A evasão no Ensino Superior Brasileiro.* Cadernos de Pesquisa, 37(132):641–659, dezembro 2007, ISSN 0100-1574. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-15742007000300007&lng=en&nrm=iso&tlng=pt, acesso em 2016-06-07. 2

[14] Aparecida, Cristiane e Baggi, Santos e Lopes Doraci Alves: *Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica.* Avaliação, 16:355–374, julho 2011. http://www.scielo.br/pdf/aval/v16n2/a07v16n2, acesso em 2016-06-08. 2

[15] Tinto, Vincent: *Leaving College: Rethinking the Causes and Cures of Student Attrition.* University of Chicago Press, Chicago, Ill.; Bristol, edição: 2nd ed. edição, 2012, ISBN 978-0-226-00757-1. 3

[16] Viana, Giomar e Lima, Jandir Ferrera de: *The human capital theory and the economic growth.* Interações, 11(2):137–148, dezembro 2010, ISSN 1518-7012. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1518-70122010000200003&lng=en&nrm=iso&tlng=pt, acesso em 2016-08-12. 3, 109

[17] Hanushek, Eric A. e Kimko, Dennis D.: *Schooling, labor-force quality, and the growth of nations.* American economic review, páginas 1184–1208, 2000. http://www.jstor.org/stable/2677847, acesso em 2016-08-12. 3

[18] Marin, Maria José Sanches e Panes, Vanessa Clivelaro Bertassi: *Envelhecimento da População e as Políticas Públicas de Saúde.* Revista do Instituto de Políticas Públicas de Marília, 1(1), 2015. http://www2.marilia.unesp.br/ojs-2.4.5/index.php/RIPPMAR/article/view/5641, acesso em 2016-07-20. 3

[19] Inep: *Sinopses da educação superior.* Disponível em: http://portal.inep.gov.br/superior-censosuperior-sinopse. 3, 111

[20] Sarker, F. e Tiropanis, T. e Davis H. C.: *Linked data, data mining and external open data for better prediction of at-risk students.* Em *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*, páginas 652–657, novembro 2014. 5, 6, 7, 19, 57

[21] Bean, John P.: *Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome.* American Educational Research Journal, 22(1):35–64, março 1985, ISSN 0002-8312, 1935-1011. `http://aer.sagepub.com/content/22/1/35`, acesso em 2016-06-07. 5

[22] Tinto, Vincent e Pusser, Brian: *Moving from theory to action: Building a model of institutional action for student success.* National Postsecondary Education Cooperative, páginas 1–51, 2006. 5

[23] Delen, Dursun: *Predicting Student Attrition with Data Mining Methods.* Journal of College Student Retention: Research, Theory & Practice, 13(1):17–35, maio 2011, ISSN 1521-0251, 1541-4167. `http://csr.sagepub.com/content/13/1/17`, acesso em 2016-06-07. 5, 6, 7, 9

[24] Noble, Kimberly e Flynn, Nicole T. e Lee James D. e Hilton David: *Predicting Successful College Experiences: Evidence from a First Year Retention Program.* Journal of College Student Retention: Research, Theory & Practice, 9(1):39–60, 2008, ISSN 1521-0251. 5, 19

[25] Dekker, Gerben W. e Pechenizkiy, Mykola e Vleeshouwers Jan M.: *Predicting Students Drop Out: A Case Study.* International Working Group on Educational Data Mining, julho 2009. `http://eric.ed.gov/?id=ED539082`, acesso em 2016-06-07. 6, 7, 57

[26] Zhang, Ying e Oussena, Samia e Clark Tony e Kim Hyeonsook: *Use Data Mining to Improve Student Retention in Higher Education - A Case Study.* Volume 1, páginas 190–197, junho 2010. 7

[27] Romero, C. e Ventura, S.: *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, novembro 2010, ISSN 1094-6977. 7, 8

[28] Arulampalam, Wiji e Naylor, Robin A. e Smith Jeremy P.: *Effects of in-class variation and student rank on the probability of withdrawal: cross-section and time-series analysis for UK university students.* Economics of Education Review, 24(3):251–262, junho 2005, ISSN 0272-7757. 7, 19, 57

[29] Meedech, Phanupong e Iam-On, Natthakan e Boongoen Tossapon: *Prediction of Student Dropout Using Personal Profile and Data Mining Approach.* Em *Intelligent and Evolutionary Systems*, número 5 em *Proceedings in Adaptation, Learning and Optimization*, páginas 143–155. Springer International Publishing, 2016, ISBN 978-3-319-26999-3 978-3-319-27000-5. DOI: 10.1007/978-3-319-27000-5_12. 7, 19, 43

[30] Langley, Pat: *The changing science of machine learning.* Machine Learning, 82(3):275–279, fevereiro 2011, ISSN 0885-6125, 1573-0565. http://link.springer.com/article/10.1007/s10994-011-5242-y, acesso em 2016-08-04. 7

[31] Witten, Ian H. e Frank, Eibe e Hall Mark A.: *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier, fevereiro 2011, ISBN 978-0-08-089036-4. 7, 8, 9, 10, 11, 12, 15, 21, 50, 51, 52, 56, 57, 59, 64

[32] Anand, Sarabjot S. e Büchner, Alex G.: *Decision support using data mining.* Financial Times Management, 1998. 8, 9

[33] Domingos, Pedro: *A Few Useful Things to Know About Machine Learning.* Commun. ACM, 55(10):78–87, outubro 2012, ISSN 0001-0782. http://doi.acm.org/10.1145/2347736.2347755, acesso em 2016-05-19. 8, 9

[34] Wu, Xindong e Kumar, Vipin e Quinlan J. Ross e Ghosh Joydeep e Yang Qiang e Motoda Hiroshi e McLachlan Geoffrey J. e Ng Angus e Liu Bing e Yu Philip S. e Zhou Zhi Hua e Steinbach Michael e Hand David J. e Steinberg Dan: *Top 10 algorithms in data mining.* Knowledge and Information Systems, 14(1):1–37, dezembro 2007, ISSN 0219-1377, 0219-3116. http://link.springer.com/article/10.1007/s10115-007-0114-2, acesso em 2016-06-12. 8, 9, 10

[35] Fayyad, Usama M. e Piatetsky-Shapiro, Gregory e Smyth Padhraic e others: *Knowledge discovery and data mining: towards a unifying framework.* Em *KDD*, volume 96, páginas 82–88, 1996. http://www.aaai.org/Papers/KDD/1996/KDD96-014, acesso em 2016-07-07. 8, 9, 16

[36] Shearer, Colin: *The CRISP-DM model: the new blueprint for data mining.* Journal of data warehousing, 5(4):13–22, 2000. 8, 16

[37] Grossman, Robert L. e Kamath, Chandrika e Kegelmeyer Philip e Kumar Vipin e Namburu Raju: *Data mining for scientific and engineering applications*, volume 2. Springer Science & Business Media, 2013. 8

[38] Carvalho, Ricardo Silva: *Modelos preditivos para avaliação de risco de corrupção de servidores públicos federais.* Dissertação, Universidade de Brasília, Brasília -DF, janeiro 2016. http://repositorio.unb.br/handle/10482/19361, acesso em 2016-08-04. 8

[39] Azevedo, Ana Isabel Rojão Lourenço e Santos, Manuel Filipe: *Kdd, semma crisp-dm: a parallel overview.* IADIS European Conference Data Mining, páginas 182–185, 2008. https://recipp.ipp.pt/handle/10400.22/135, acesso em 2016-07-07. 8

[40] Kurgan, Lukasz A. e Musilek, Petr: *A survey of Knowledge Discovery and Data Mining process models.* The Knowledge Engineering Review, 21(1):1–24, março 2006, ISSN 1469-8005, 0269-8889. https://www.cambridge.org/core/journals/knowledge-engineering-review/article/a-survey-of-knowledge-discovery-and-data-mining-process-models/368D6AFE435EB5E30378398D34D61C17, acesso em 2017-11-17TZ. 9

[41] Cios, Krzysztof J. e Swiniarski, Roman W. e Pedrycz Witold e Kurgan Lukasz A.: *The knowledge discovery process*. Em *Data Mining*, páginas 9–24. Springer, 2007. http://link.springer.com/content/pdf/10.1007/978-0-387-36795-8_2.pdf, acesso em 2016-07-07. 9, 16

[42] Cabena, Peter e Hadjinian, Pablo e Stadler Rolf e Verhees Jaap e Zanasi Alessandro: *Discovering Data Mining: From Concept to Implementation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998, ISBN 978-0-13-743980-5. 9, 16

[43] Chapman, Pete e Clinton, Julian e Kerber Randy e Khabaza Thomas e Reinartz Thomas e Shearer Colin e Wirth Rudiger: *CRISP-DM 1.0 Step-by-step data mining guide*. 2000. http://www.citeulike.org/group/1598/article/1025172, acesso em 2016-08-04. 9, 16, 17

[44] Hosmer Jr, David W. e Lemeshow, Stanley: *Applied logistic regression*. John Wiley & Sons, 2004. 9, 54

[45] Hastie, Trevor e Tibshirani, Robert e Friedman Jerome: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009, ISBN 978-0-387-84857-0 978-0-387-84858-7. http://link.springer.com/10.1007/978-0-387-84858-7, acesso em 2016-06-07. 10, 13, 14, 50, 51, 52, 59

[46] Quinlan, J. Ross: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993, ISBN 978-1-55860-238-0. 11, 12, 55

[47] Quinlan, J. Ross: *Ross Quinlan personal webpage*. http://www.rulequest.com/Personal/, acesso em 2016-06-30TZ. 11

[48] Kuhn, Max: *C50 package documentation*. https://cran.r-project.org/web/packages/C50/C50.pdf, acesso em 2016-06-30TZ. 11

[49] Quinlan, J. Ross: *Comparing C4.5 and C5.0*. http://rulequest.com/see5-comparison.html, acesso em 2016-06-30TZ. 11

[50] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. http://www.R-project.org, ISBN 3-900051-07-0. 11

[51] Breiman, Leo e Friedman, Jerome e Stone Charles J. e Olshen R. A.: *Classification and Regression Trees*. Chapman and Hall/CRC, New York, 1 edition edição, janeiro 1984, ISBN 978-0-412-04841-8. 12, 54, 70

[52] Domingos, Pedro e Pazzani, Michael: *On the optimality of the simple Bayesian classifier under zero-one loss*. Machine learning, 29(2-3):103–130, 1997. http://link.springer.com/article/10.1023/A:1007413511361, acesso em 2016-08-07. 12, 14, 15, 64

[53] Haykin, Simon: *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2a edição, 1998, ISBN 978-0-13-273350-2. 12, 13

[54] EUROSTAT: *Classificação internacional eurostat/unesco/ocde*, outubro 2000. http://download.inep.gov.br/download/superior/2009/Tabela_OCDE_2009.pdf, acesso em 2016-01-08. 17, 23

[55] Ishitani, Terry T.: *A Longitudinal Approach to Assessing Attrition Behavior Among First-Generation Students: Time-Varying Effects of Pre-College Characteristics*. Research in Higher Education, 44(4):433–449, agosto 2003, ISSN 0361-0365, 1573-188X. 19, 43

[56] Boero, G. e Laureti, T. e Naylor R.: *An econometric analysis of student withdrawal and progression in post-reform Italian Universities*. Working Paper CRENoS 200504, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia, 2005. 19

[57] Webber, Douglas A. e Ehrenberg, Ronald G.: *Do expenditures other than instructional expenditures affect graduation and persistence rates in American higher education?* Economics of Education Review, 29(6):947–958, 2010. 19

[58] Zimmermann, Judith e Brodersen, Kay H. e Heinimann Hans R. e Buhmann Joachim M.: *A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance*. JEDM - Journal of Educational Data Mining, 7(3):151–176, outubro 2015, ISSN 2157-2100. 19

[59] Inep: *Manual de Preenchimento do censo da educação superior | 2016 - Módulo aluno*, fevereiro 2017. http://download.inep.gov.br/educacao_superior/censo_superior/questionarios_e_manuais/2016/manual_aluno2016.pdf, acesso em 2017-05-27. 20, 21, 28, 33

[60] Brasil: *Portaria nº 40/MEC. Institui o e-MEC, sistema eletrônico de fluxo de trabalho e gerenciamento de informações relativas aos processos de regulação da educação superior no sistema federal de educação.*, dezembro 2007. http://download.inep.gov.br/download/condicoes_ensino/2007/Portaria_n40.pdf, acesso em 2016-06-15. 21

[61] Brasil: *Portaria nº 438/MEC. Institui o Exame Nacional do Ensino Médio - ENEM*, maio 1998. http://www.crmariocovas.sp.gov.br/pdf/diretrizes_p0178-0181_c.pdf, acesso em 2017-05-16. 23, 24

[62] Brasil: *Portaria nº 80/MEC. Institui o Exame Nacional do Ensino Médio - ENEM*, junho 2010. https://www.legisweb.com.br/legislacao/?id=227492, acesso em 2017-05-16. 23, 24

[63] Brasil: *Portaria nº 21/MEC. Institui o Sistema de Seleção Unificada - SISU*, maio 2012. http://static07.mec.gov.br/sisu/portal/data/portaria.pdf, acesso em 2016-06-15. 23, 24

[64] Brasil: *Lei nº 11.096. Institui o Programa Universidade para Todos - PROUNI.*, janeiro 2005. http://siteprouni.mec.gov.br/doc/Portaria%20normativa%206%20de%2026%20de%20fevereiro%202014%20-%20atualizada.pdf, acesso em 2017-05-16. 24

[65] Brasil: *Portaria nº 170/MEC. Dispões sobre a ocupação de vagas do processo seletivo do FIES.*, abril 2017. http://fiesselecao.mec.gov.br/arquivos/portaria_normativa_16_01092017.pdf, acesso em 2017-09-14. 24

[66] Ahmed, Abeer Badr El Din e Elaraby, Ibrahim Sayed: *Data Mining: A prediction for Student's Performance Using Classification Method.* World Journal of Computer Application and Technology, 2(2):43–47, fevereiro 2014. http://www.hrpub.org/journals/article_info.php?aid=1285, acesso em 2017-09-18TZ. 39

[67] Kabakchieva, Dorina: *Predicting Student Performance by Using Data Mining Methods for Classification.* Cybernetics and Information Technologies, 13(1):61–72, 2013, ISSN 1314-4081. https://www.degruyter.com/view/j/cait.2013.13.issue-1/cait-2013-0006/cait-2013-0006.xml, acesso em 2017-09-18. 39

[68] He, H. e Garcia, E. A.: *Learning from Imbalanced Data.* IEEE Transactions on Knowledge and Data Engineering, 21(9):1263–1284, setembro 2009, ISSN 1041-4347. 51

[69] John, George H. e Langley, Pat: *Estimating Continuous Distributions in Bayesian Classifiers.* Em *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, páginas 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc., ISBN 978-1-55860-385-1. http://dl.acm.org/citation.cfm?id=2074158.2074196, acesso em 2017-11-02TZ. 54

[70] Bussab, Wilton de O. e Morettin, Pedro A.: *Estatística básica*, volume 1. Saraiva, São Paulo, 2010. 59, 78

[71] Garson, G.D.: *Interpreting neural network connection weights.* Artificial Intelligence Expert, 6(4):46–51, 1991. 64

Apêndice A

Resultados de desempenho durante treinamento dos modelos analisados

Tabela A.1: Evasão a nível de área de estudo – Desempenho médio dos modelos de *Naive Bayes* nas *tabelas* de treinamento obtido através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

——	Kernel	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	Não	78,97%	42,35%	81,28%
	Sim	92,20%	10,19%	97,38%
Downsampling	Não	62,58%	60,50%	64,65%
	Sim	62,52%	53,92%	71,13%
Upsampling	Não	64,73%	60,20%	69,26%
	Sim	65,61%	57,93%	73,28%

Tabela A.2: Evasão a nível de área de estudo – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	94,07%	0,00%	100,00%
1	0,0001	94,07%	0,00%	100,00%
1	0,1000	94,07%	0,00%	100,00%
3	0,0000	94,07%	0,00%	100,00%
3	0,0001	94,07%	0,00%	100,00%
3	0,1000	92,28%	12,52%	97,31%
5	0,0000	94,07%	0,00%	100,00%
5	0,0001	94,06%	0,00%	99,99%
5	0,1000	91,13%	18,89%	95,69%

Tabela A.3: Evasão a nível de área de estudo – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	56,10%	72,10%	40,08%
1	0,0001	52,07%	39,19%	65,00%
1	0,1000	60,77%	57,75%	63,81%
3	0,0000	54,89%	54,47%	55,36%
3	0,0001	54,78%	76,11%	33,44%
3	0,1000	62,68%	67,72%	57,63%
5	0,0000	53,45%	82,57%	24,32%
5	0,0001	59,50%	65,92%	53,08%
5	0,1000	61,14%	62,32%	59,96%

Tabela A.4: Evasão a nível de área de estudo – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	49,99%	25,01%	74,97%
1	0,0001	53,77%	52,30%	55,23%
1	0,1000	64,60%	71,24%	57,97%
3	0,0000	58,66%	48,62%	68,71%
3	0,0001	68,27%	54,18%	82,37%
3	0,1000	85,46%	91,41%	79,52%
5	0,0000	63,73%	44,14%	83,31%
5	0,0001	74,97%	76,62%	73,32%
5	0,1000	86,66%	89,69%	83,63%

Tabela A.5: Evasão a nível de área de estudo – Desempenho médio dos modelos de regressão logística nas *tabelas* de treinamento obtidos através da validação cruzada tamanho 4 por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	93,58%	6,79%	99,06%
Downsampling	64,86%	63,27%	66,46%
Upsampling	76,75%	75,10%	78,39%

Tabela A.6: Evasão a nível de IES – Desempenho médio dos modelos de *Naive Bayes* nas *tabelas* de treinamento obtido através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

——	Kernel	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	Não	80,05%	44,55%	81,96%
	Sim	93,01%	10,34%	97,47%
Downsampling	Não	63,04%	58,42%	67,66%
	Sim	62,93%	58,75%	67,10%
Upsampling	Não	64,82%	60,62%	69,03%
	Sim	65,78%	56,58%	74,98%

Tabela A.7: Evasão a nível de IES – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	94,88%	0,00%	100,00%
1	0,0001	94,88%	0,00%	100,00%
1	0,1000	94,61%	6,04%	99,39%
3	0,0000	94,88%	0,00%	100,00%
3	0,0001	94,88%	0,00%	100,00%
3	0,1000	92,81%	18,16%	96,84%
5	0,0000	94,88%	0,00%	100,00%
5	0,0001	94,88%	0,00%	100,00%
5	0,1000	91,65%	16,72%	95,69%

Tabela A.8: Evasão a nível de IES – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	55,56%	51,32%	59,85%
1	0,0001	56,05%	66,79%	45,30%
1	0,1000	56,16%	44,10%	68,22%
3	0,0000	56,32%	44,66%	68,00%
3	0,0001	58,75%	52,80%	64,69%
3	0,1000	64,57%	69,30%	59,85%
5	0,0000	58,36%	43,23%	73,50%
5	0,0001	56,77%	73,02%	40,46%
5	0,1000	65,01%	64,46%	65,56%

Tabela A.9: Evasão a nível de IES – Desempenho médio dos modelos de redes neurais na *tabela* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração.

Nós	Decaimento	Acurácia	Sensibilidade	Especificidade
1	0,0000	53,20%	38,15%	68,26%
1	0,0001	55,48%	64,40%	46,57%
1	0,1000	59,95%	59,10%	60,81%
3	0,0000	69,53%	51,23%	87,84%
3	0,0001	61,81%	54,72%	68,90%
3	0,1000	86,50%	88,57%	84,44%
5	0,0000	62,89%	66,24%	59,56%
5	0,0001	59,80%	79,20%	40,39%
5	0,1000	87,76%	92,70%	82,82%

Tabela A.10: Evasão a nível de IES – Desempenho médio dos modelos de regressão logística nas *tabelas* de treinamento obtidos através da validação cruzada tamanho 4 por tipo de balanceamento.

——	Acurácia	Sensibilidade	Especificidade
Sem balanceamento	94,58%	8,36%	99,22%
Downsampling	65,90%	65,01%	66,78%
Upsampling	78,16%	79,73%	76,59%

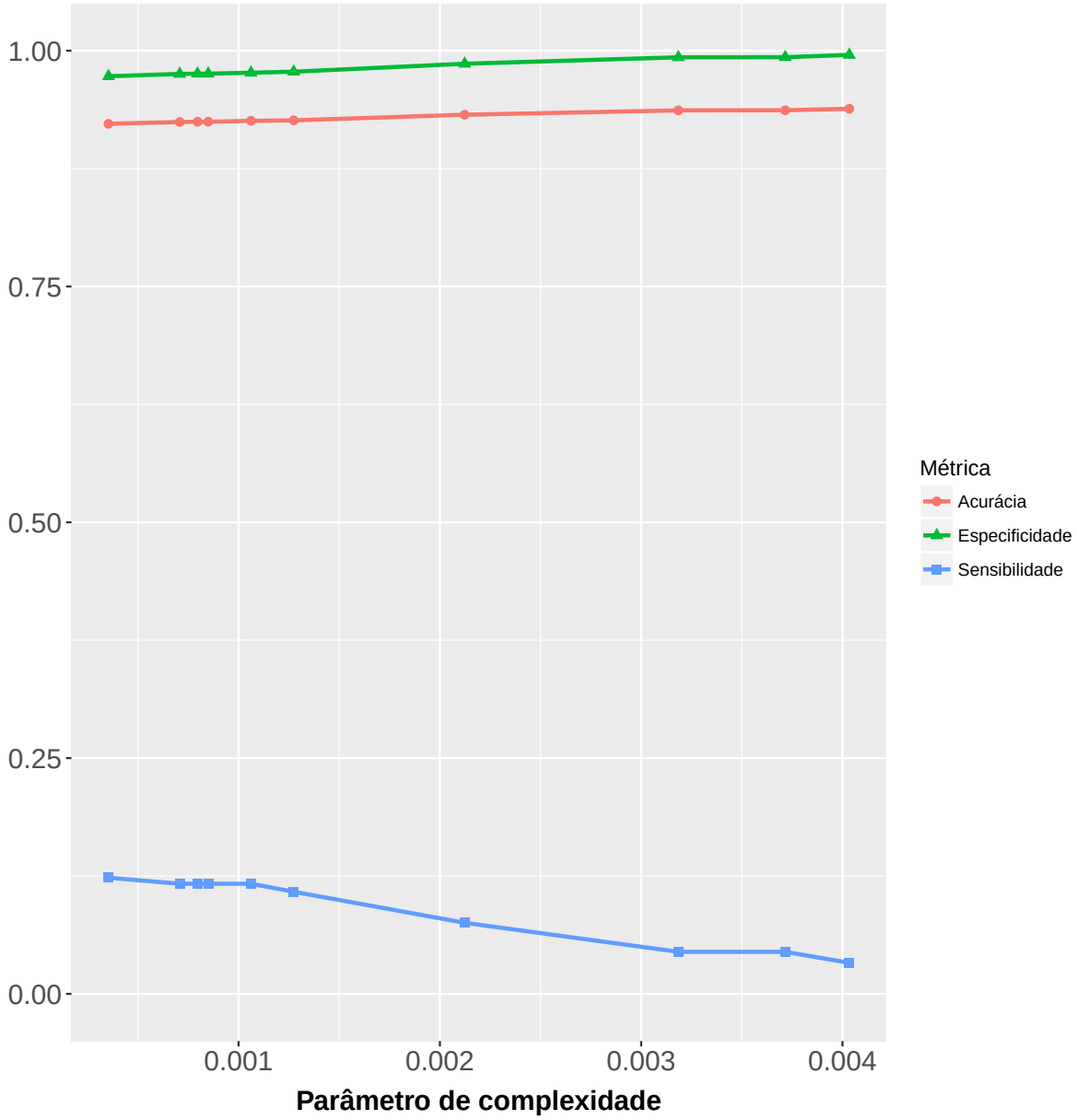

Figura A.1: Evasão a nível de área de estudo – Desempenho médio dos modelos CART nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

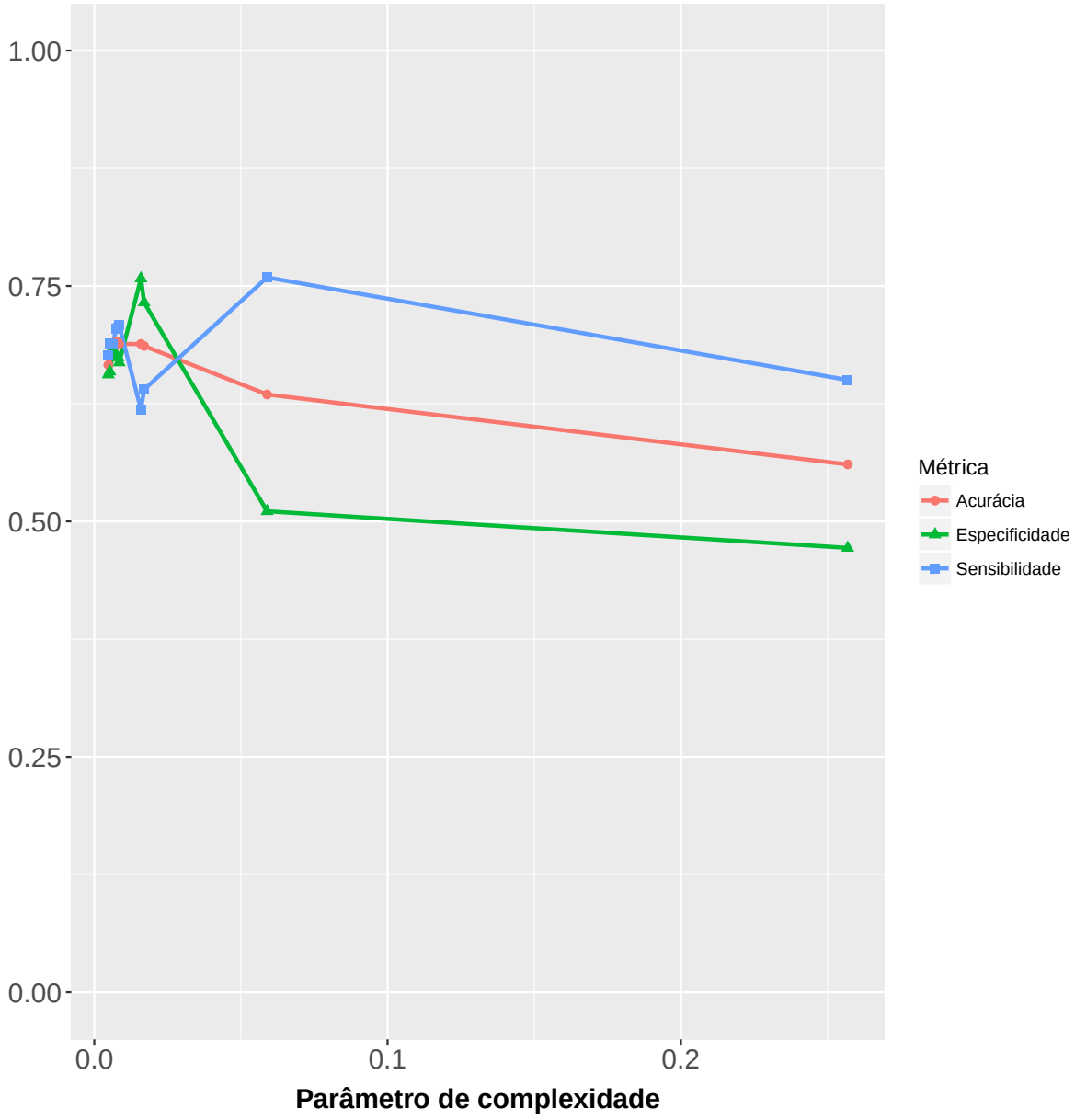

Figura A.2: Evasão a nível de área de estudo – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

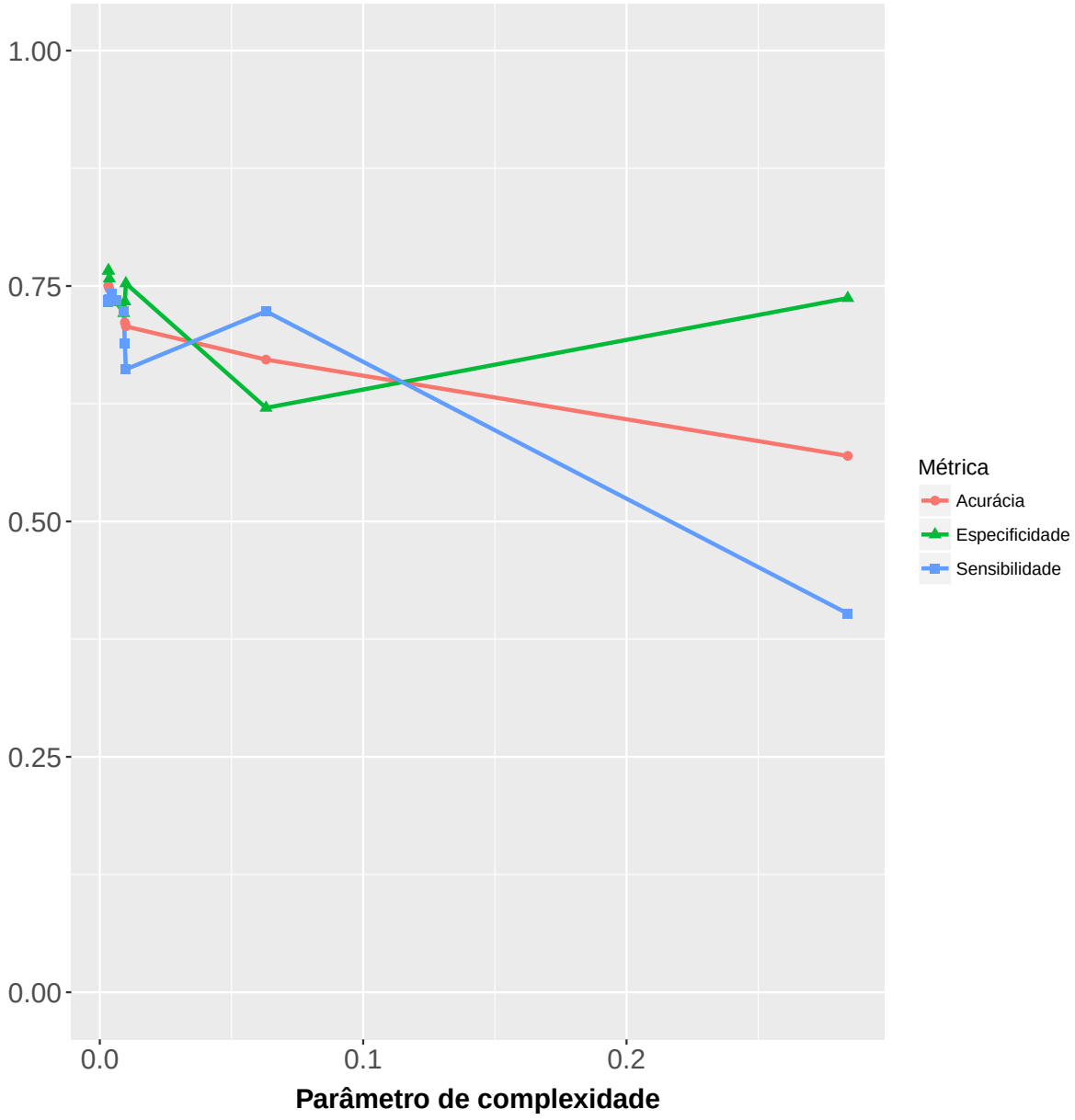

Figura A.3: Evasão a nível de área de estudo – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

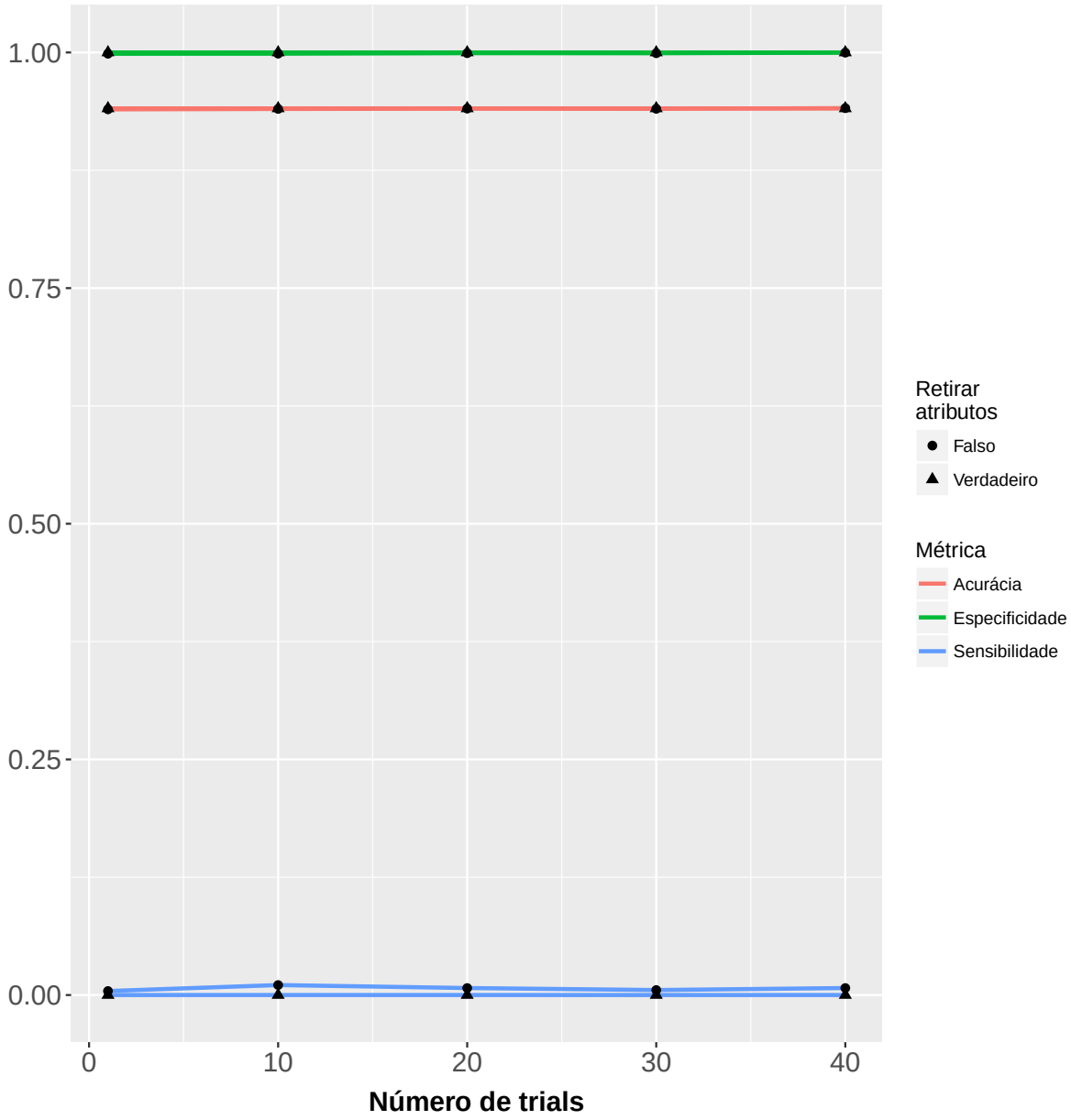

Figura A.4: Evasão a nível de área de estudo – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

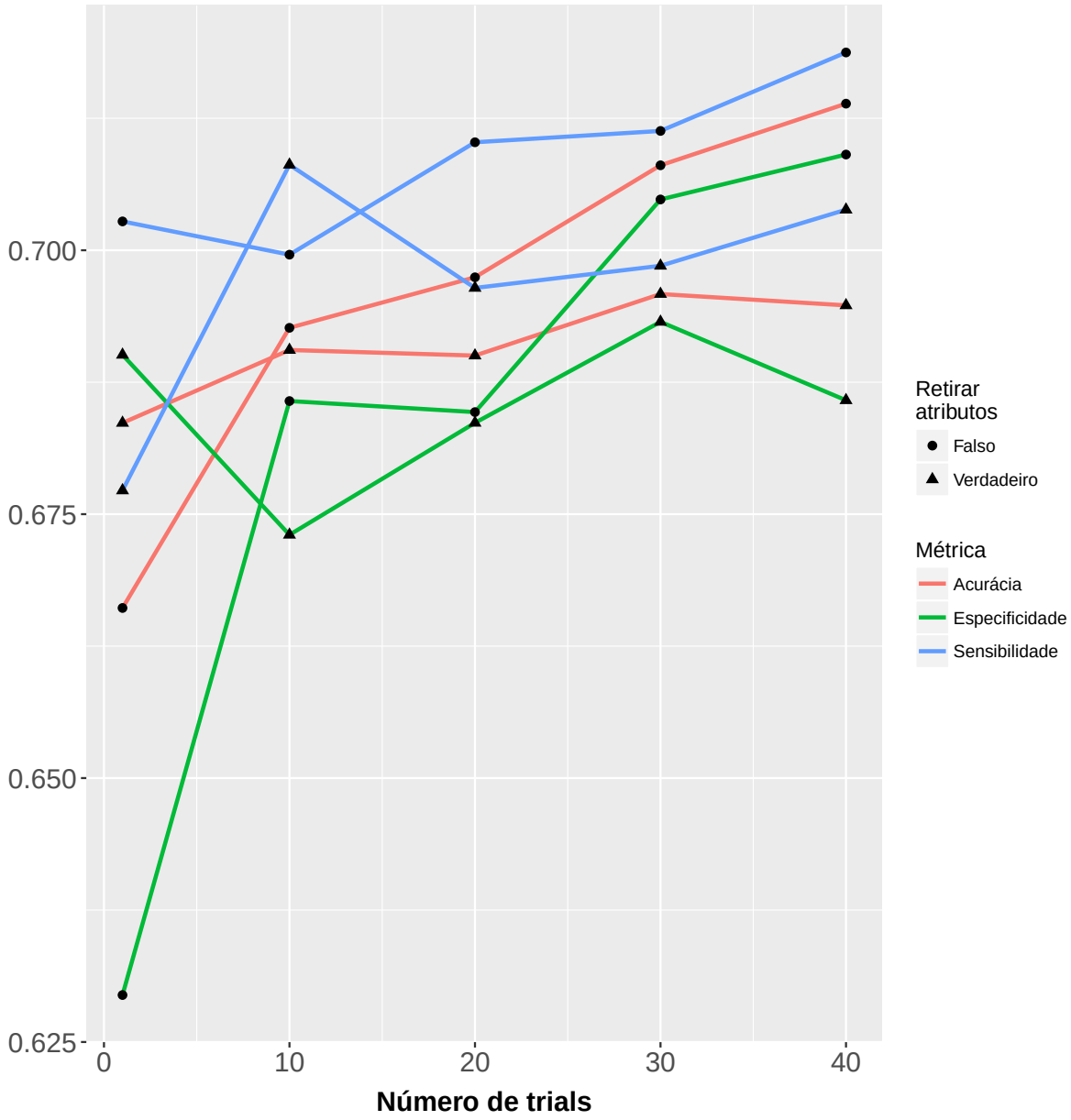

Figura A.5: Evasão a nível de área de estudo – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

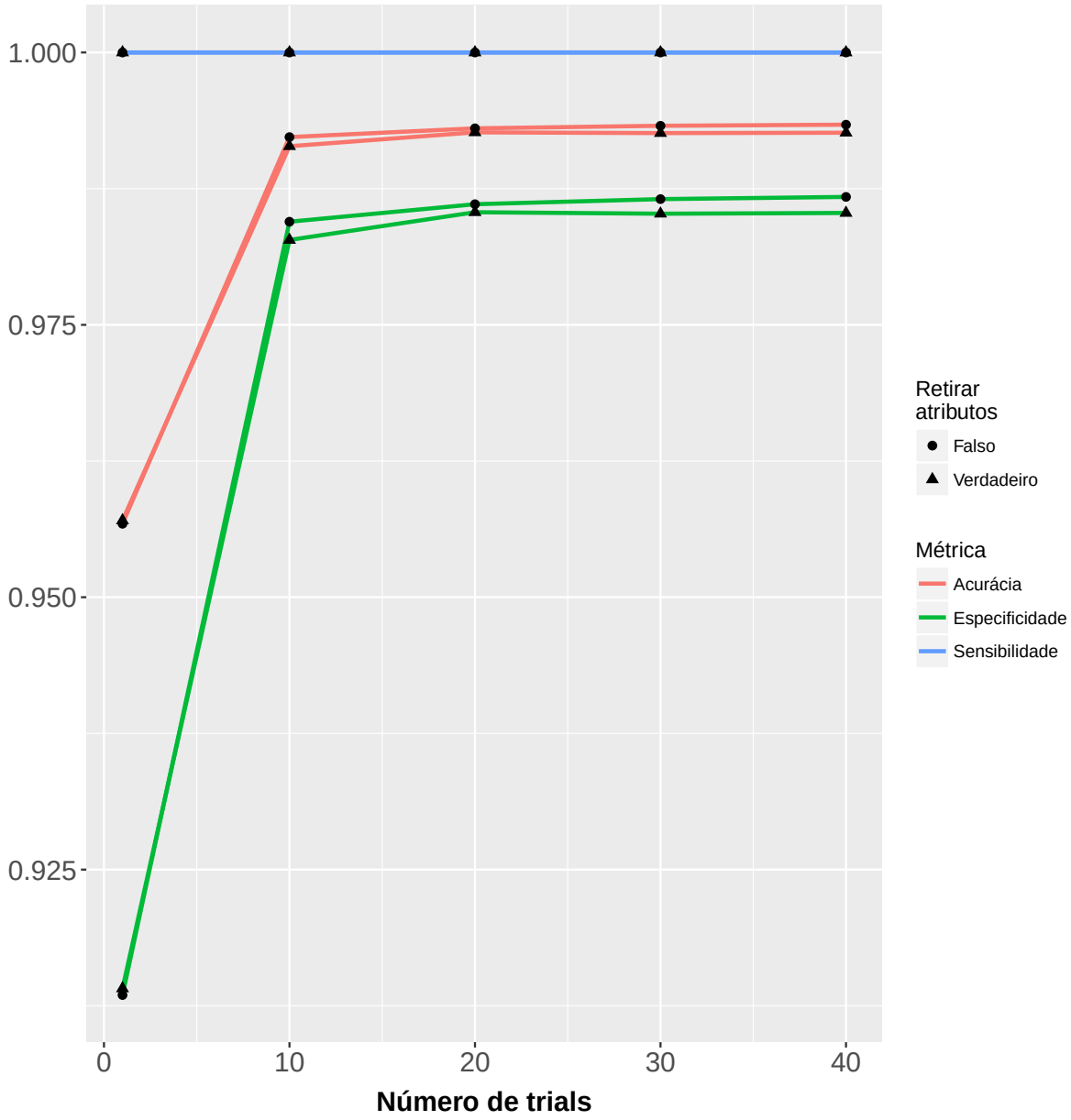

Figura A.6: Evasão a nível de área de estudo – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

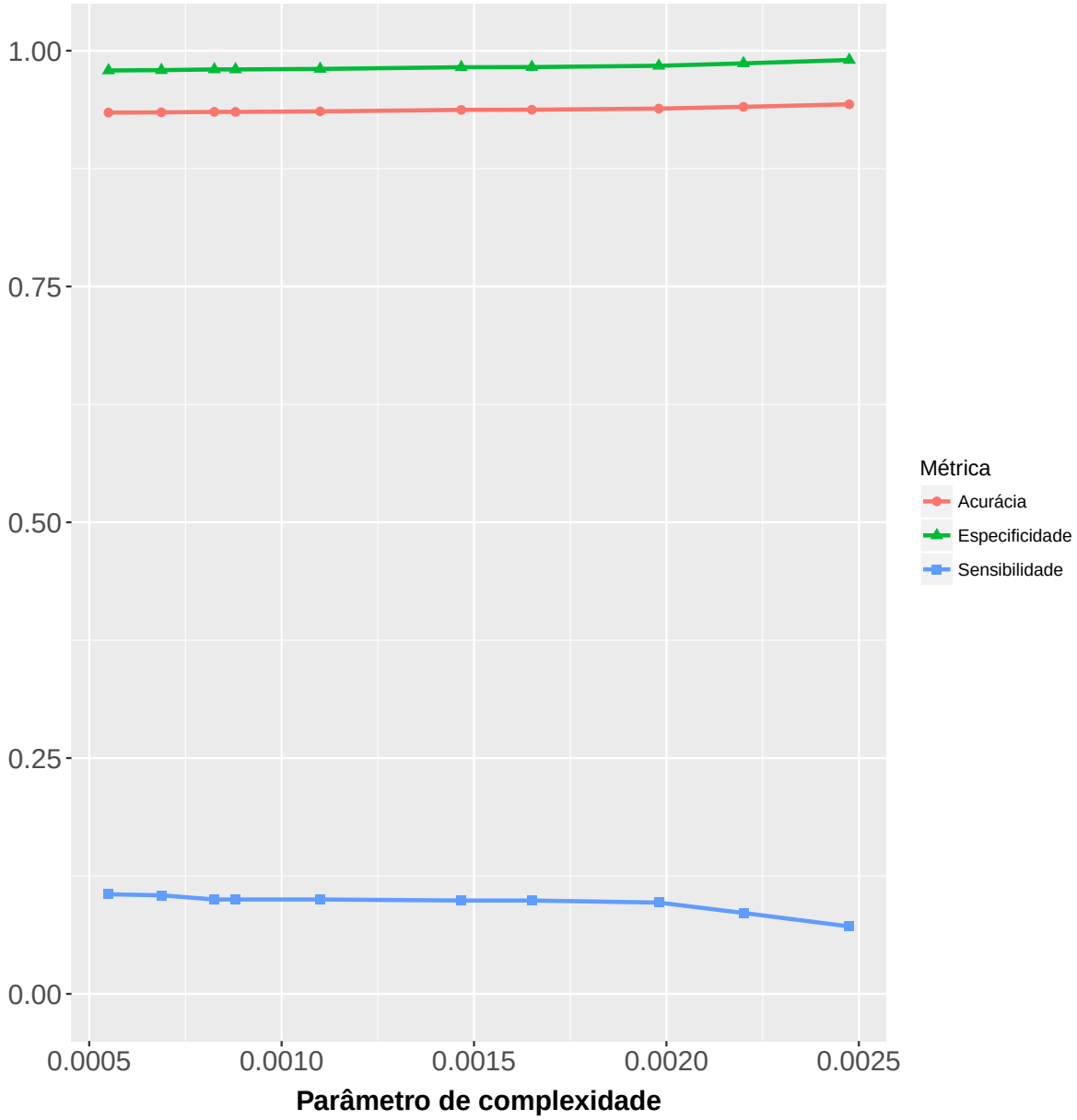

Figura A.7: Evasão a nível de IES – Desempenho médio dos modelos CART nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

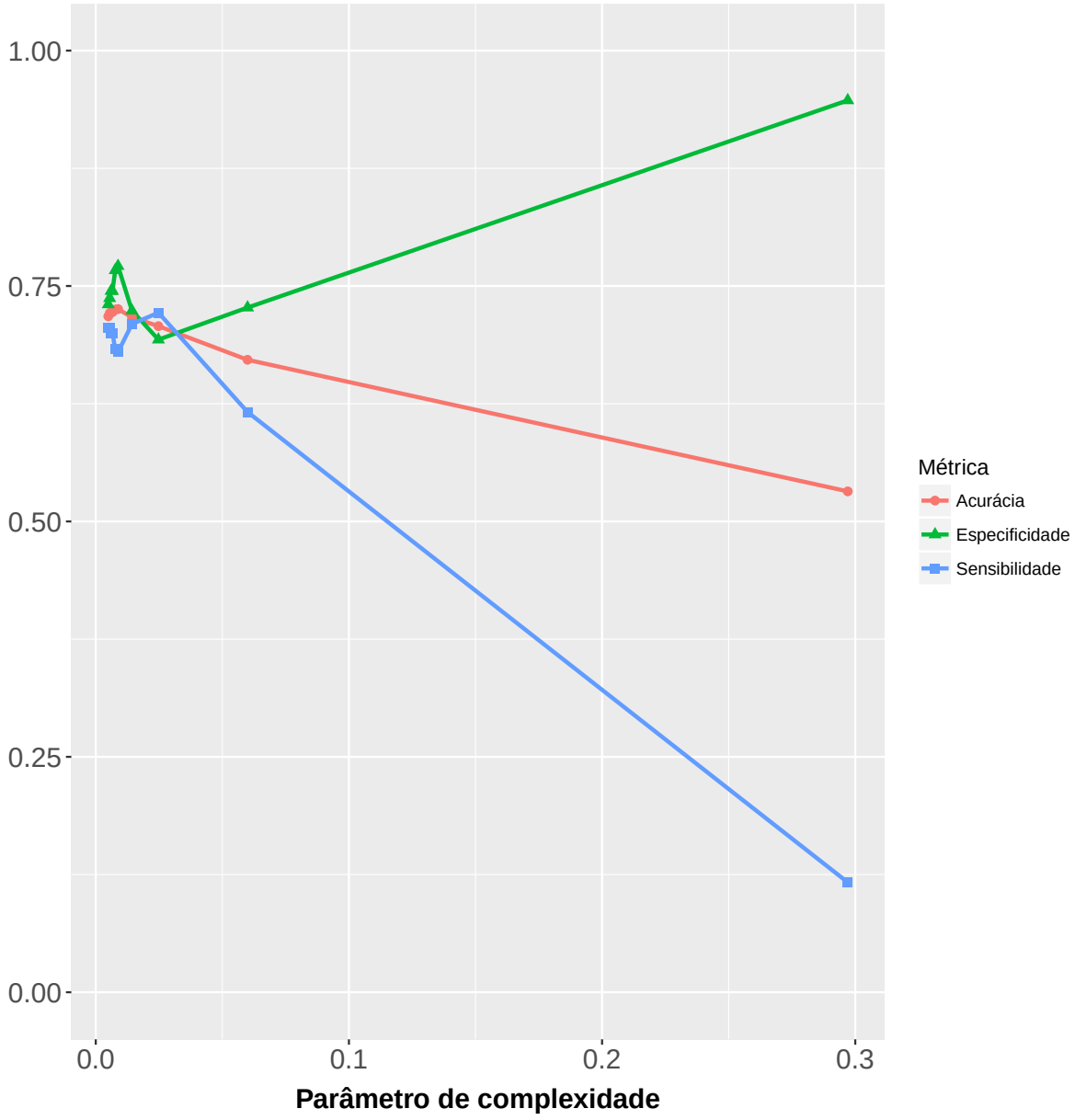

Figura A.8: Evasão a nível de IES – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

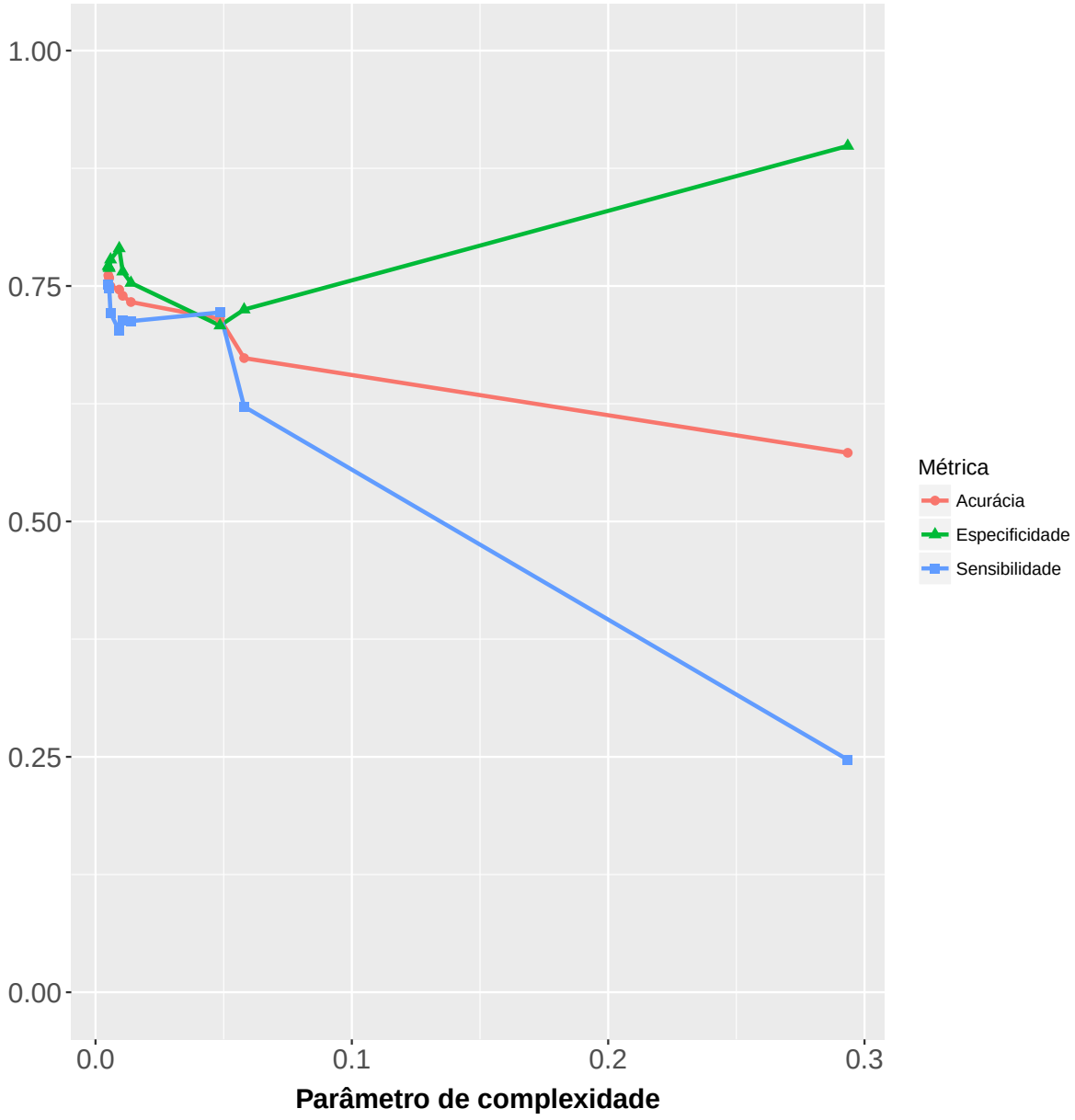

Figura A.9: Evasão a nível de IES – Desempenho médio dos modelos CART nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

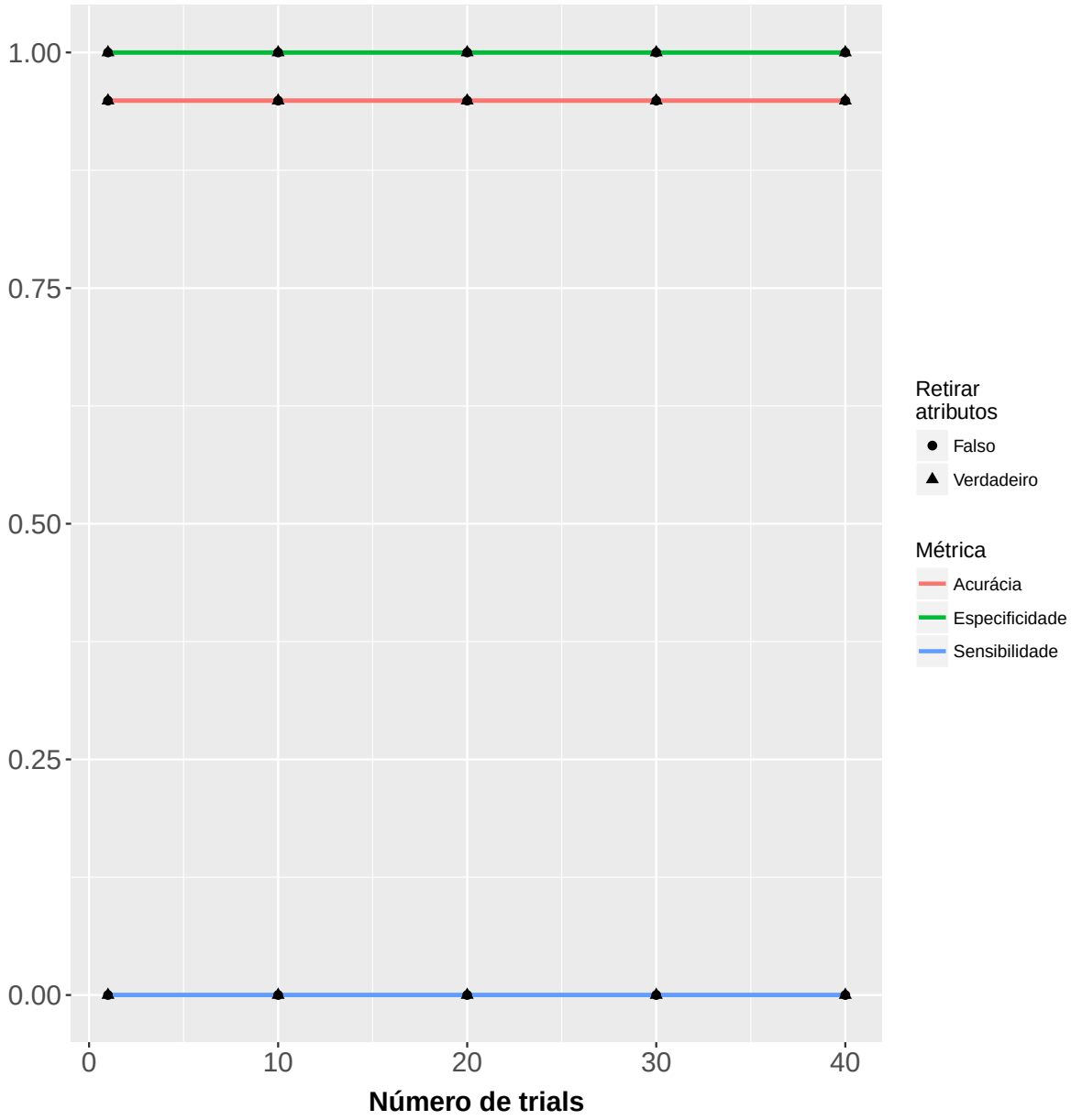

Figura A.10: Evasão a nível de IES – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento **sem balanceamento** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

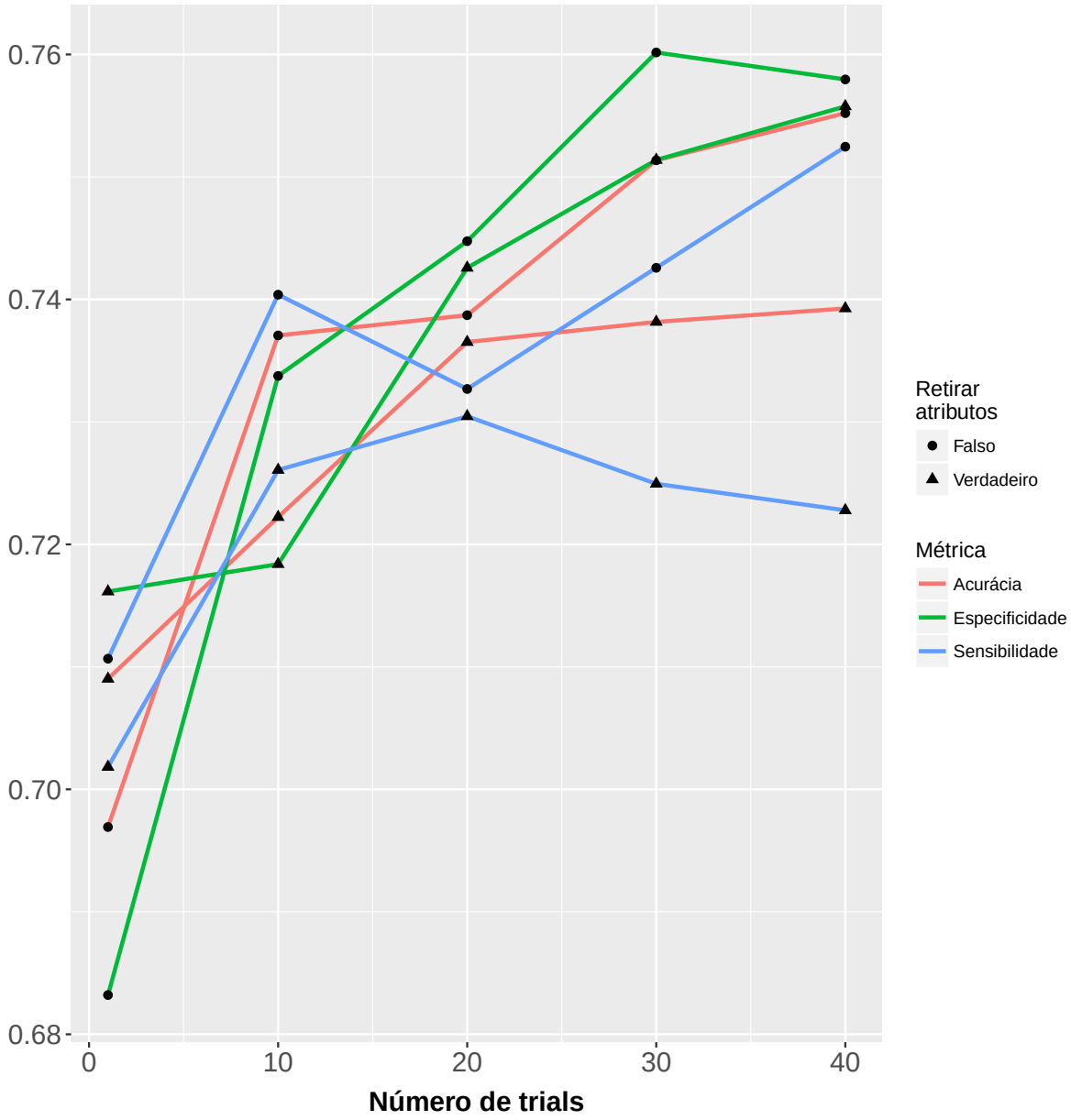

Figura A.11: Evasão a nível de IES – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***downsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.

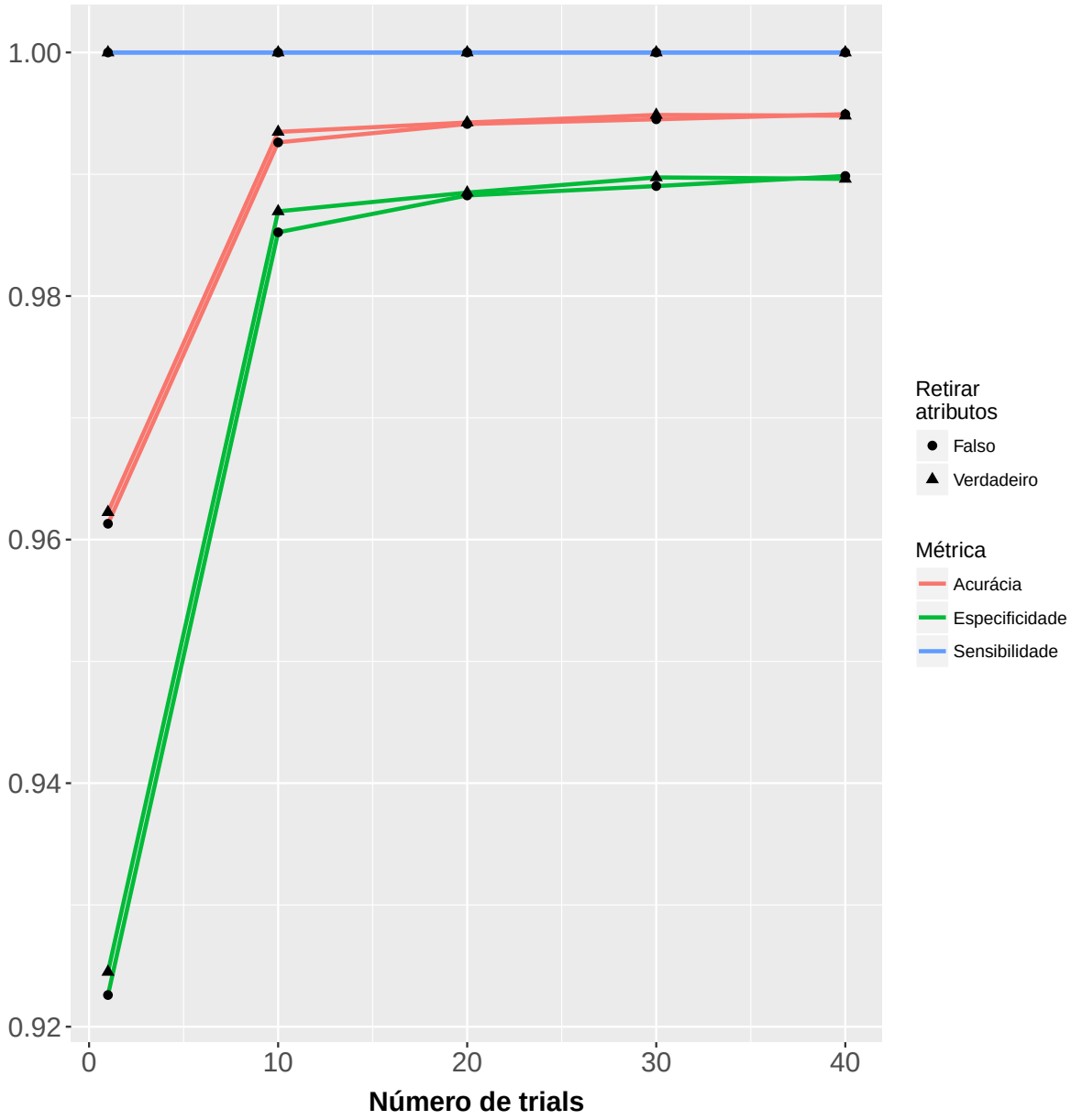

Figura A.12: Evasão a nível de IES – Desempenho médio dos modelos C5.0 nas *tabelas* de treinamento com ***upsampling*** obtidos através da validação cruzada tamanho 4 por tipo de configuração e balanceamento.