PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Francisco Coimbra Carneiro Pereira

Modelos Preditivos para Evasão de Alunos no Ensino Superior Privado – Uma aplicação de *Machine Learning* para Gestão de Marketing de Relacionamento

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Empresas da PUC-Rio como requisito parcial para obtenção do grau de Mestre em Administração de Empresas.

Orientador: Prof. Jorge Brantes Ferreira

Rio de Janeiro
Abril de 2017

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Francisco Coimbra Carneiro Pereira

Modelos Preditivos para Evasão de Alunos no Ensino Superior Privado – Uma aplicação de *Machine Learning* para Gestão de Marketing de Relacionamento

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Administração de Empresas da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Jorge Brantes Ferreira
Orientador
Departamento de Administração – PUC-Rio

Prof.ª Angela Maria Cavalcanti da Rocha
Departamento de Administração - PUC-Rio

Prof. Marcus Wilcox Hemais
Departamento de Administração - PUC-Rio
(Suplente)

Prof. Renato Dourado Cotta de Mello
UFRJ

Profª. Mônica Herz
Coordenadora Setorial do Centro
de Ciências Sociais – PUC-Rio

Rio de Janeiro, 12 de abril de 2017

Francisco Coimbra Carneiro Pereira

Graduou-se em Ciências Econômicas pela Faculdade de Economia e Finanças - IBMEC/RJ, em 2012.

Ficha Catalográfica

Pereira, Francisco Coimbra Carneiro

Modelos preditivos para evasão de alunos no ensino superior privado : uma aplicação de *machine learning* para gestão de marketing de relacionamento / Francisco Coimbra Carneiro Pereira ; orientador: Jorge Brantes Ferreira. – 2017.

69 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Administração, 2017.

Inclui bibliografia

1. Administração – Teses. 2. Machine learning. 3. Modelos preditivos. 4. Classificação. 5. Evasão de alunos. 6. Lifetime value. I. Ferreira, Jorge Brantes. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Administração. III. Título.

CDD: 658

Agradecimentos

Ao meu orientador, professor Jorge Brantes pelo seu apoio e paciência durante este percurso acadêmico, compreendendo minhas limitações, mas sempre atencioso e disposto a ajudar.

A Alexandre Mathias, pela inspiração e atração pelo tema da educação superior.

À Comatrix Gestão e Consultoria e aos sócios Claudio Zohar, Marcel Sapir e Marcio Roza que me apoiaram e ajudaram a cursar este mestrado.

A Rodolfo Bertolini e João Carlos Padilha Silva pela ajuda e informações compartilhadas.

Aos meus pais Letícia e Ricardo, pela educação que me deram.

Aos meus irmãos Filipe e Carolina.

Resumo

Pereira, Francisco Coimbra Carneiro; Ferreira, Jorge Brantes. **Modelos Preditivos para Evasão de Alunos no Ensino Superior Privado - Uma aplicação de Machine Learning para Gestão de Marketing de Relacionamento**. Rio de Janeiro, 2017. 69p. Dissertação de Mestrado - Departamento de Administração, Pontifícia Universidade Católica do Rio de Janeiro.

Perdendo em média mais de 20% da base de alunos todo semestre, a evasão de alunos no ensino superior privado representa um desafio para a gestão dessas instituições. Diferentes abordagens são utilizadas para combater este problema. Para a gestão de marketing de retenção, a identificação dos alunos é o primeiro passo necessário para aplicar uma estratégia de interação personalizada. Nesse sentido, este trabalho apresenta uma metodologia quantitativa para classificação de risco de evasão de alunos ativos. Baseado em dados históricos de alunos que evadiram ou se formaram, modelos gerados por algoritmos de *machine learning* foram calculados e comparados e, na sequência, utilizados para classificar alunos ativos. Por fim, estimou-se o *lifetime value* desses alunos para auxiliar na definição de estratégias de retenção.

Palavras- chave

Machine learning, modelos preditivos, classificação, evasão de alunos, lifetime value.

Abstract

Pereira, Francisco Coimbra Carneiro; Ferreira, Jorge Brantes. (Advisor). **Predictive Models for Student Attrition in Private Graduation – An application of Machine Learning to Relationship Marketing Management.** Rio de Janeiro, 2017. 69p. Dissertação de Mestrado - Departamento de Administração, Pontifícia Universidade Católica do Rio de Janeiro.

Losing more than 20% of its students each semester, the student attrition in private graduation courses challenges its institutions management. Different approaches to address this problem have been used. To retention marketing management the identification of students is the first necessary step to apply a personalized interaction strategy. In this sense, this work uses a quantitative methodology to classify its students by risk of attrition. Based in historic data of former students of an institution, models were generated by machine learning algorithms and its results compared. Then they were used to classify active students in the educational institution. Afterwards, their lifetime value were estimated in order to help in the definition of retention strategies.

Keywords

Machine learning, predictive models, classification, student attrition, lifetime value.

Sumário

Lista de figuras

Lista de tabelas

1.
O problema

O número de alunos matriculados no ensino superior brasileiro cresceu de 3,8 milhões em 2003, para 7,3 milhões de alunos em 2013 (INEP, 2015), com a expansão da oferta ocorrendo, principalmente, no ensino privado e, em particular, em instituições posicionadas para atender às classes C e D. O aumento do poder aquisitivo da população brasileira na década de 2000, os programas federais de bolsas (ProUni) e de financiamento subsidiado (FIES) para o estudo superior em instituições privadas, além de fatores do mercado de trabalho, como a distância salarial existente entre portadores de diploma de ensino médio vis-à-vis de ensino superior foram alguns fatores importantes para alavancar o crescimento desse mercado. No entanto, tão expressiva quanto a curva de crescimento da base é a curva de evasão, que mesmo com o crescimento da demanda, manteve-se acima de 20% por semestre ao longo do período. Esse é um problema sobretudo no ensino privado, onde existência da variável preço apresenta-se como um elemento de *trade-off* na decisão do aluno de permanecer na instituição.

O conceito de evasão pode ser abrangente, incluindo três tipos: a evasão do curso, da instituição e do sistema (LOBO, 2012). Na primeira, o aluno muda de curso, mas permanece na Instituição de Ensino Superior (IES), enquanto na segunda, muda de IES. A terceira evasão, do sistema, representa o abandono do aluno do ensino superior. Do ponto de vista da IES, as duas últimas representam uma perda já que, na mudança de curso, o aluno continua na instituição, apenas consumindo um serviço diferente.

Numa abordagem de *Customer Relationship Management* (CRM), uma base de alunos com menor taxa de evasão representa uma base de clientes mais valiosa, pois estes possuem um maior *lifetime value* para a instituição. Ainda, quanto maior a taxa de evasão, maior o esforço de captação necessário para manter o nível da base. Além disso, um indicador chave para a eficiência de uma IES, o número de alunos por turma – que permite a diluição do custo docente – é prejudicado por uma evasão mais alta, pois torna-se mais difícil compor turmas maiores para disciplinas dos períodos finais. Diferentemente do ensino básico, o ingresso no ensino superior ocorre, basicamente, no primeiro ano do curso, de modo

que a base que evade ao longo dos períodos não é reposta, levando a um menor número de alunos por turma em períodos mais avançados e, consequentemente, a uma menor margem de contribuição dessas turmas. Portanto, a capacidade de retenção não apenas aumenta o valor econômico das IES por meio de um maior número de alunos gerando receita, como também por meio do aumento da rentabilidade da operação. Tendo em vista o impacto da taxa de evasão no valor econômico das IES (SCHWARTZMAN, 2003), a compreensão das suas causas torna-se vital para que as instituições possam agir preventivamente para manter os alunos na sua base.

Mas quais são, afinal, as causas de taxas de evasão tão altas no ensino superior brasileiro? No Brasil, estudos recentes que têm pesquisado esse fenômeno buscam explica-lo a partir de fatores intrínsecos às instituições e cursos (SILVA FILHO et al., 2007), aos alunos (BORGES, 2011) ou na relação entre esses dois atores (PEREIRA JÚNIOR, 2012). Algumas relações importantes foram encontradas, como a menor taxa de evasão em cursos com maior concorrência na entrada (SILVA FILHO et al., 2007). Porém, frequentemente nos casos em que houve pesquisa quantitativa, a coleta de dados ocorreu ou por meio de questionários com alunos, que por dificuldades próprias do método, representam amostras reduzidas em relação ao universo total de estudantes nas IES, ou por meio de dados agregados das instituições. Será que existem padrões invisíveis nos dados tradicionalmente utilizados que podem contribuir para explicar a evasão? As instituições possuem sistemas informatizados que armazenam grandes quantidades de dados sobre cada aluno, como todas as notas que obteve em cada curso, frequência, distância de sua residência à instituição, número de boletos pagos em atraso etc., porém a dificuldade de acesso de pesquisadores a esses bancos de dados impede que esses sejam usados em estudos sobre o tema, ainda que as próprias IES não sejam capazes de utilizar todo o potencial dessas informações. Portanto, a possibilidade de acessar esses tipos de dados, internos à instituição, representa uma oportunidade rara de aplicação da metodologia escolhida para este estudo em contexto acadêmico no Brasil, com possíveis desdobramentos não apenas para a instituição alvo escolhida, mas também para outras IES quanto ao potencial valor implícito em suas bases de dados.

1.1. Objetivos

Este estudo tem como propósito a resolução de um problema gerencial de uma IES privada, e apresenta objetivos intermediários relacionados à aplicação de metodologia de *machine learning* para identificação e diferenciação de alunos.

1.1.1. Objetivo Final

O objetivo deste estudo é estimar o risco de evasão de alunos de uma IES privada a partir de dados acadêmicos e financeiros. Para isso, o estudo se propõe a, por meio de técnicas de *machine learning* utilizando informações disponíveis nos bancos de dados da IES, desenvolver um modelo preditivo para a evasão a partir de subconjuntos de variáveis dessas dimensões. Com isso, o objetivo final é que a IES possa ter uma ferramenta para atuar preventivamente na retenção de alunos.

1.1.2. Objetivos Intermediários

1. Coletar, consolidar e preparar num só banco de dados as informações relevantes para o problema, tendo como fontes os dados dos alunos presentes nos sistemas utilizados pela IES.

2. Aplicar técnicas de *machine leargning* (árvores de decisão e regressão logística) no banco de dados unificado para encontrar modelos que expliquem a taxa de evasão a partir das variáveis disponíveis.

3. Analisar os modelos e variáveis relevantes e refinar a explicação da variável alvo (evasão).

4. Estimar a probabilidade de evasão de uma amostra de alunos ativos na instituição a partir dos modelos desenvolvidos.

5. Apresentar impactos teóricos e gerenciais advindos dos resultados do estudo.

1.2.
Delimitação do Estudo

O estudo será realizado com dados secundários, extraídos de bancos de dados históricos de alunos matriculados em uma instituição de ensino superior privada do município do Rio de Janeiro, com posicionamento para a classe C. A amostra contém alunos que foram matriculados em cursos de graduação oferecidos pela IES desde 2010.

1.3.
Justificativa e Relevância do Estudo

A computação trouxe, segundo o especialista em CRM Stan Rapp, três grandes capacidades: as capacidades de armazenar, de encontrar e de comparar informações (RAPP; COLLINS, 1995). Com o avanço exponencial da tecnologia da informação e, consequentemente, dessas capacidades, surge o conceito de *Big Data* e o campo de estudo em marketing de CRM. Em alguns setores de forma mais proeminente do que em outros, a utilização de grande quantidade de dados sobre clientes para melhor entendê-los e atingi-los é uma realidade nas organizações.

No setor de educação, no entanto, apesar do grande volume de dados registrados sobre seus alunos, essas informações ainda são subutilizadas pelas instituições de ensino brasileiras. Muitas instituições já incorporam algumas ferramentas de CRM, principalmente para direcionar mensagens de email marketing e ligações, porém tais ações, além de, em geral, tratarem os alunos de maneira pouco individualizada, representam apenas uma pequena fração do potencial de valor que essas bases de informação podem gerar às IES.

Do ponto de vista de CRM, um dos *drivers* principais para aumentar o valor dos clientes é aumentar a sua retenção – ao prolongar o tempo de relacionamento com o cliente, aumenta-se o número de transações com este, elevando seu valor para empresa. Aplicado ao ensino superior, é possível aumentar o valor da base de alunos para a IES ao reduzir a taxa de evasão. Para isso, é preciso entender quais são os fatores que levam um aluno a evadir para que a IES possa agir preventivamente. Embora haja na literatura no Brasil estudos que investigam as razões da evasão de alunos no ensino superior (ADACHI 2009, LOBO 2012, PRIM e FÁVERO 2013, GERBA 2014), essas pesquisas abordam a questão de maneira

qualitativa ou com pouco aprofundamento estatístico devido a limitações de acesso a dados. A relevância desta pesquisa está, precisamente, no acesso exclusivo aos dados internos de uma IES, permitindo a utilização métodos de *data mining* utilizando um conjunto amplo de dados secundários, disponíveis nos próprios sistemas utilizados pela IES, envolvendo características socioeconômicas, acadêmicas e do relacionamento financeiro dos estudantes com a instituição, abrangendo milhares de alunos e sua evolução ao longo do tempo.

A compreensão dos fatores para a evasão é particularmente relevante para as IES na conjuntura atual da economia brasileira, com inflação e desemprego em alta e redução da oferta de crédito pelo FIES, combinada à característica de maior sensibilidade a preço dos estudantes de baixa renda em instituições privadas. Uma nova abordagem investigativa pode contribuir para que novas ferramentas sejam utilizadas para retenção de alunos no ensino superior privado.

2.
Referencial Teórico

Este estudo pretende abordar a evasão estudantil no ensino superior privado como um problema de gestão de marketing. Os conceitos teóricos utilizados baseiam-se em CRM e *machine learning*, que têm como função identificar clientes para que esses possam ser tratados de forma individualizada, aumentando a eficácia das ações de marketing.

2.1.
O Conceito de CRM (Customer Relationship Management)

A evolução dos mercados e das empresas no século XX mudou o modelo de atuação comercial das empresas, de pequenos negócios locais que conheciam cada um de seus clientes e suas necessidades, para grandes negócios globais onde a comunicação para as massas tornou-se o principal e mais eficaz meio para atingir os consumidores. Todavia a evolução tecnológica trouxe um novo paradigma de marketing para a forma como se dá a relação das empresas com os clientes. A disponibilidade massiva de dados e a capacidade de processamento computacional possibilitam novamente a individualização de cada cliente. Há uma mudança de foco de grandes segmentos de mercado para consumidores individuais, com as empresas buscando identificá-los individualmente, seus hábitos e desejos.

Nesse cenário surge o CRM – *Customer Relationship Management*, área de marketing que tem como atividade aprender mais sobre as necessidades e comportamentos dos clientes para desenvolver melhores estratégias de relacionamento com a empresa (BOLTON, 1998). O objetivo final é aumentar o valor da base de clientes da empresa, ao torná-los mais fiéis por meio de maiores níveis de satisfação proporcionados por uma melhor experiência relacional. O aprendizado sobre as características dos clientes é feito a partir de informações coletadas e produzidas ao longo do tempo. Nas interações dos clientes com as empresas, a depender do tipo de ambiente de negócios, diversas informações podem ser produzidas e armazenadas em bancos de dados. Essas informações podem ser utilizadas para (i) identificar, (ii) diferenciar, (iii) interagir e (iv) customizar ações individualizadas para cada cliente. Quanto mais personalizadas forem essas ações, maior será sua eficácia em atingir os interesses e necessidades do cliente

(PEPPERS; ROGERS, 2011). O CRM encoraja uma visão de organização baseada na gestão de clientes, em oposição a uma gestão de produtos. Ao invés de cruzar dados para identificar potenciais clientes na base atual para um novo produto, busca-se, muito mais, identificar qual a necessidade de cada cliente, para, a partir desse conhecimento, oferecer o produto que melhor o atende. Há evidências empíricas na literatura (PEPPERS & ROGERS GROUP, 2000) da relação entre o nível de utilização de ferramentas de CRM com a lealdade de longo prazo do cliente com a empresa. O cliente reconhece valor no tratamento individualizado que ele recebe. Para Barnes (2001), o aumento do valor percebido pelo cliente em cada interação com a empresa aumenta seu nível de satisfação, levando a taxas maiores de retenção. Thakur e Summey (2010) ilustram alguns resultados que podem ser alcançados via processos de CRM: (1) prover melhor serviço ao cliente; (2) ajudar a área comercial a fechar vendas mais rápido; (3) melhorar a eficácia de *cross-sellings*; (4) simplificar processos de marketing e comerciais; (5) descobrir novos clientes, (6) fazer *call centers* mais eficientes e; (7) aumentar a receita por cliente.

Conceitualmente e de uma maneira mais abrangente, pode-se dizer que o marketing de relacionamento visa três objetivos finais: atrair novos clientes, reter a base de clientes e aumentar o valor de sua base de clientes. Esta tríade é conhecida como Get-Keep-Grow (PEPPERS; ROGERS, 2011). Desses objetivos, particularmente a retenção (Keep) é o objetivo em termos de CRM para o qual este trabalho está focado. Retenção significa manter um cliente atual comprando da mesma empresa, em detrimento de produtos ou serviços concorrentes, ou do abandono do consumo da categoria de produto como um todo. Há muitas evidências empíricas de que a retenção, em negócios diversos, é mais lucrativa – ou, vista de outro ângulo, menos custosa – do que a aquisição de novos clientes (KOTLER, 1994; HOGAN *et al.*, 2003; LEE-KELLEY *et al.*, 2003; BLATTBERG; DEIGHTON, 1996; FILIATRAUL; LAPIERRE, 1997). A manutenção de clientes ao longo do tempo está estritamente ligada ao conceito de lealdade.

2.1.1. Lealdade

Lealdade é um conceito central no marketing de relacionamento. O resultado de uma boa gestão de relacionamento com o cliente é um maior nível de

lealdade deste com a instituição. Para Dick e Basu (1994) a lealdade do consumidor é dada pela força da relação entre a atitude do indivíduo com um produto, marca ou empresa e a compra repetitiva da mesma. Assim, a lealdade requer tanto elementos psicológicos afetivos e cognitivos que formam a atitude, como elementos comportamentais, expressos pela recorrência da compra. Nesse sentido, lealdade se diferencia de simples retenção, que requer apenas o elemento comportamental de recompra, mas não necessariamente o atitudinal. Embora essa interação entre atitude e comportamento apareçam recorrentemente em diversos autores (ROWLEY; DAWES, 2000; OLIVER, 1999), há divergências na academia quanto à exata relação dessas variáveis no processo de formação de lealdade (BERGAMO e GIULIANI, 2009).

Uma estratégia de CRM bem sucedida é capaz de reter o cliente pois este foi convertido num cliente leal, cujo comportamento de compra é complementado por uma atitude bastante positiva da empresa. Dessa forma, não apenas a manutenção (Keep) é garantida, mas também probabilidade de este cliente comprar mais no futuro (Grow).

2.2.
Marketing e retenção no setor de educação superior

2.2.1.
Evasão de alunos no ensino superior

Vincent Tinto (1975, 1987, 1993) é frequentemente citado por autores diversos sobre evasão e, embora antigo, apresenta em 1975 um modelo para compreensão sobre o tema que se tornou referência na área. Este modelo foi revisitado pelo autor novamente em 1987 e 1993, complementando seu arcabouço teórico. Para analisar o problema da evasão, Tinto recorre a elementos de psicologia e baseia-se na teoria do suicídio de Durkheim (1961) – para este, o suicídio é mais provável de ocorrer em indivíduos não suficientemente integrados à sociedade, particularmente quando faltam ao indivíduo integração moral (seus valores não se alinham aos da sociedade) e afiliação coletiva (não se sente parte de uma organização maior). Analogamente, Tinto trata o campus acadêmico como um sistema social com seus próprios valores e estruturas, analisando o abandono da

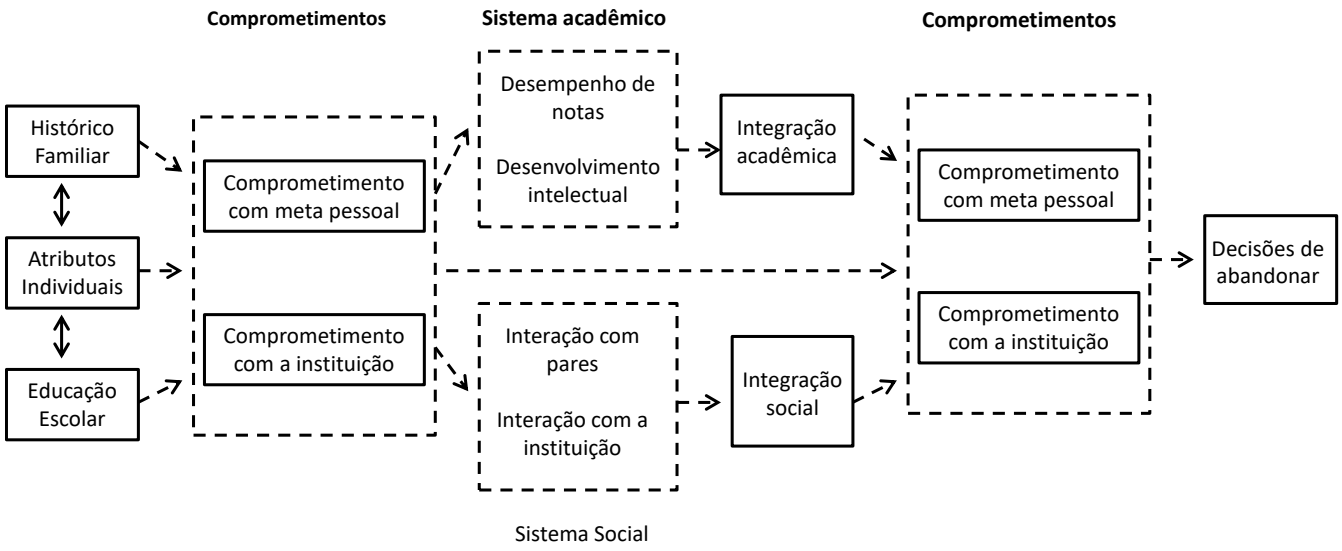
instituição de ensino como o suicídio para Durkheim. Sob esta ótica, argumenta que uma insuficiência de integração social com outros habitantes desse microcosmos e a ausência de alinhamento de valores do aluno com a instituição são fatores que contribuem para a evasão.

No entanto, além de um sistema social, a instituição de ensino também é um sistema acadêmico e, embora perfeitamente integrado no sistema social, o aluno pode falhar no sistema acadêmico – por baixo desempenho escolar – e ser desligado da instituição. O inverso também é possível, com um estudante, aparentemente bem sucedido academicamente, evadindo da instituição. O autor aponta uma relação entre essas duas esferas: até certo ponto, pode haver uma relação positiva entre ambas, com um aluno com bom desempenho tendo maior aceitação no sistema social e sentindo-se mais alinhado à instituição, ou o aluno que, bem relacionado dentro desse sistema, torna-se mais engajado na esfera acadêmica. Todavia, há um limite para essa complementariedade positiva, tomando como exemplo o aluno que se engaja excessivamente em atividades sociais em detrimento de seu tempo de estudo.

Portanto, para Tinto (1975) é preciso distinguir os tipos de evasão, particularmente decisões espontâneas ou falha no desempenho acadêmico, pois esses tipos envolvem diferentes tipos de estudante e diferentes padrões de relacionamento com a instituição. Enquanto para o segundo grupo, uma falha de aderência ao sistema acadêmico é mais evidente, no primeiro grupo, parece haver menor integração social e aproximação ideológica com a instituição. Para o autor, instituições de ensino maiores tendem a reduzir a distância ideológica e falta de integração do aluno, pois há espaço para mais subestruturas, com maior diversidade e, potencialmente, maior chance de o aluno se conectar a algumas dessas. Outros fatores de distinção entre esses dois grupos são o status social e o nível intelectual. No caso dos alunos com fraco desempenho acadêmico, tanto o nível intelectual quanto o status social tendem a ser mais baixos, enquanto que os alunos que evadem por decisão própria tendem a demostrar exatamente o oposto. Nesse sentido, programas da instituição voltados a ajudar o desempenho acadêmico de alunos provenientes de famílias de mais baixa renda, por exemplo, fazem sentido para aumentar a retenção na visão do autor.

No modelo de Tinto (1975), dois conceitos são chave, a integração (aos sistemas social e acadêmico) e o comprometimento. Esses se relacionam num processo dinâmico, apresentado na figura 2.1.

Figura 2.1. – Estrutura do modelo de decisão de evasão de Tinto (1975)



Fonte: Tinto (1975)

Fatores antecedentes ao ingresso na instituição, quais sejam, o histórico familiar, atributos individuais e a educação escolar, moldam o comprometimento do aluno com suas próprias metas e com a instituição. O nível de comprometimento com suas metas influencia a integração ao sistema acadêmico, enquanto o comprometimento com a instituição influencia a integração ao sistema social. Todavia, o processo se mostra dinâmico na medida em que o nível de integração a esses sistemas também afeta o nível de comprometimento futuro do aluno. Eventualmente, se um dos dois níveis de comprometimento se mostrar muito baixo, o aluno abandona o curso. Para Tinto (1975), o comprometimento do aluno com seus objetivos e com a instituição tem influência direta na lealdade. Em 1993, o autor introduz uma terceira variável de comprometimento, o comprometimento com atividades externas, tendo relação negativa com a lealdade.

2.2.2.
Aproximação de marketing e retenção de alunos

Os trabalhos de Tinto (1975, 1987, 1993) marcaram a literatura sobre o tema na área de educação, não obstante críticas de alguns autores (BEAN; METZNER 1985, GRUBB 1989, TIERNEY 1992, SEIDMAN 1996), porém uma maior aproximação de sua teoria sobre evasão com marketing de relacionamento é um fenômeno mais recente. Inspirados pelo modelo de Tinto, Hennig-Thurau, Langer e Hansen (2001) apresentam um modelo de lealdade baseado na qualidade da relação aluno-instituição (batizado de *relationship quality-based student loyalty*, referenciado como RQSL). Para eles, a principal crítica ao modelo é o papel excessivamente central do constructo comprometimento em detrimento de outros fatores. Por exemplo, para Tinto (1975) a qualidade dos professores é um determinante para o nível de integração do aluno, mas não uma variável de impacto direto na lealdade do aluno em si. Tinto (1975) coloca as mudanças comportamentais dos alunos como determinantes da lealdade, enquanto mudanças na instituição ou outros elementos da prestação do serviço educacional em si não são considerados diretamente.

Hennig-Thurau, Langer e Hansen (2001) integram o modelo de Tinto ao conceito de qualidade da relação, dando a instituição um papel mais central na habilidade de gerar níveis mais altos de lealdade. No modelo RQSL a lealdade é função três constructos: a percepção dos estudantes da qualidade da educação (a qualidade do serviço), a confiança do estudante nos atores da instituição (funcionários acadêmicos ou administrativos) e o comprometimento do aluno com a instituição. Após testes empíricos com alunos de universidades públicas na Alemanha, os autores identificaram na percepção de qualidade a mais forte relação positiva com lealdade, seguido do comprometimento emocional do aluno com a instituição, sendo o primeiro constructo duas vezes mais relevante que o segundo.

2.3.
O Conceito de Margem de Contribuição no Setor de Educação

Margem de contribuição é uma linha de resultado de contabilidade gerencial em que busca-se separar os gastos não em custos (ligados diretamente a atividade fim) e despesas (atividades meio), mas sim em custos fixos e variáveis. Entende-se

por custo variável todos os gastos cujo montante em um determinado período é função direta do volume de produção, em oposição ao custo fixo que, embora também tenha natureza flutuante, esta não oscila diretamente em função do volume no curto prazo – evidentemente, em longo prazo, a estrutura de custo fixo é dimensionada para um determinado volume de produção esperado, de modo a minimizar o custo total médio (custo total / volume). A margem de contribuição tem algumas características que a tornam um indicador relevante para a gestão de um negócio. Uma delas é avaliar o impacto de uma venda adicional no resultado da empresa: a margem de contribuição elimina a necessidade de rateios de custos fixos, que podem mascarar a verdadeira contribuição econômica que uma venda adicional traz para empresa. Outra finalidade é poder calcular o ponto de equilíbrio: quantas vendas são necessárias para pagar o custo fixo e manter a empresa com lucro zero – com isso, qualquer volume de vendas superior a esse garantirá lucro positivo.

Em educação, o *output* não é um produto, e sim um serviço, por isso a métrica de volume apropriada é o número de alunos atendidos em um dado período de tempo. O custo mais relevante para uma instituição de ensino é o custo docente e este é um custo variável, embora não exatamente em função do número de alunos, mas sim, do número de turmas, esta sim, função direta da base de alunos matriculada. Podem existir outros custos diretamente associados ao número de alunos a depender da instituição (custos com envio de apostilas a alunos ou custos para acesso a conteúdo de EAD quando terceirizado, por exemplo) mas de uma maneira geral, no setor de educação utiliza-se o conceito margem de contribuição como a receita líquida menos o custo docente (LUCHESA; MACHADO, 2012). Esta é uma medida de eficiência usada para compara diferentes empresas do setor. Ela será tanto mais alta quanto, tudo mais constante, maior for o ticket médio (faturamento médio por aluno), o número de alunos por turma (maior razão entre vetor de faturamento e vetor de custo) e menor for o custo docente por hora.

2.4. *Lifetime Value e Customer Equity*

Lifetime value é uma métrica utilizada para calcular o valor de uma base de clientes ((BERGER; NASR, 1998; JAIN; SINGH, 2002). De forma genérica, seu cálculo consiste em projetar o lucro por cliente – todas as receitas menos todos os

custos diretamente relacionados ao atendimento daquele cliente – num determinado período de tempo. Os lucros futuros são então descontados a valor presente por uma taxa de juros que reflita o custo de oportunidade da empresa, analogamente ao cálculo de valor presente líquido (VPL), utilizado no campo das finanças. A fórmula a seguir expressa esse conceito (MAHISHI, 2014):

$$LV_x = \sum_{i=1}^{n} \frac{P_{x_i}}{(1+r)^i}$$

Onde x representa um cliente em particular, LV_x representa seu *lifetime value*, n representa o número de períodos considerados de relacionamento futuro com a empresa, P_{x_i} representa o lucro obtido com o cliente x no período i e r representa o custo de oportunidade da empresa. Embora o nome induza a uma ideia de valor ao longo de toda a vida do cliente, o horizonte de projeção em geral tem um prazo determinado e depende de características do negócio. Para Mahishi (2014) o horizonte de projeção depende tipicamente de três fatores: (i) competitividade da indústria – quanto mais alta, menor tende a ser a lealdade dos clientes (assim como a lucratividade) e, portanto, menor o horizonte; (ii) natureza do produto ou serviço – para compras contratuais espera-se um período de *lifetime value* ao menos igual ao tempo do contrato, enquanto para não-contratuais, fatores como a frequência de compras influenciam na duração esperada da relação; (iii) estágio da vida do consumidor – tomando como exemplo o caso de um curso superior, tudo mais constante, um aluno no 2º período terá um *lifetime value* superior a outro no 7º, simplesmente porque o segundo está no fim do ciclo esperado dentro do modelo de negócio.

Ou seja, pela ótica do *lifetime value* um cliente vale o fluxo de caixa líquido que ele gera para a empresa – todas as receitas menos todos os custos associados aos esforços para entregar o produto ou serviço. O valor total de uma base de clientes é, portanto, o somatório de todos esses *lifetime values* individuais. Há algumas dificuldades práticas em sua aplicação (VOGEL; EVANSCHITZKY; RAMASESHAN, 2008), pois embora faça sentido conceitualmente, a definição do horizonte de projeção adequado e as estimativas de lucros futuros abrem espaço para diferentes visões. Não é objetivo deste estudo discutir metodologias para

cálculo do *lifetime value*, todavia para fins de cálculo neste estudo, associou-se o conceito de lucro do cliente ao de margem de contribuição. Já para o horizonte de projeção, considerou-se o número de semestres até a expectativa de formatura, mas ao final, ponderou-se *o lifetime value* obtido por uma variável de probabilidade de conclusão, estimada a partir de modelos de *machine learning*. Pela característica do setor, a relação cliente-empresa tem um prazo determinado (pelo número de semestres do curso), mas não é exatamente um contrato, com o cliente podendo sair a qualquer momento. Por isso, optou-se por ponderar o *lifetime value* total (considerando a plenitude do que seria o prazo contratual) por uma probabilidade de o cliente ficar até o final do contrato. Assim, o *lifetime value* de um aluno na graduação é a margem de contribuição desse aluno pelo tempo esperado até sua formatura, trazida a valor presente, multiplicada pela probabilidade de conclusão.

O somatório dos *lifetime values* representa o *customer equity*, o valor daquela base de clientes para a empresa. O *customer equity* também pode ser estimado a partir de pesquisas com consumidor, levando em consideração elementos mais intangíveis como a lealdade do cliente (VOGEL; EVANSCHITZKY; RAMASESHAN, 2008). Muitas vezes esse é o caminho adotado devido às próprias dificuldades de se obter o *lifetime value*, mas uma vez que este já foi calculado, somá-los para obter uma estimativa de *customer equity* é um caminho natural.

2.5. Os Conceitos de Data Mining e Machine Learning

Data mining, ou mineração de dados, é a atividade de analisar grandes quantidades de dados com o intuito de encontrar padrões e relações significativas (BERRY; LINOFF, 2000) entre as variáveis. Existem dois tipos de *data mining*: o direto e o indireto. O *data mining* direto tenta explicar ou categorizar uma variável alvo definida e normalmente é utilizada para modelagem preditiva. O *data mining* indireto, por sua vez, busca por padrões ou similaridades entre grupos de dados de uma base sem a definição de uma variável alvo particular. Essas abordagens não são mutuamente excludentes e, na verdade, frequentemente as atividades de *data mining* utilizam as duas.

Em geral, a mineração de dados envolve ao menos uma das seguintes atividades: classificação, estimação e predição – atividades de mineração direta; agrupamento por afinidades, *clustering*, descrição e visualização – mineração indireta (HAN; KAMBER, 2006). Junto com essas atividades, diferentes algoritmos podem ser utilizados, como árvores de decisão, redes neurais e sistemas bayesianos. Diferentemente de métodos tradicionais de trabalho com dados, em que um modelo é descrito pelo pesquisador e, na sequência, seus parâmetros são estimados, esses métodos geram múltiplos modelos e os testa automaticamente.

Estas mesmas técnicas são aplicadas no domínio do que é conhecido como *machine learning*[1] e muitas vezes essas terminologias se confundem. No entanto, pode-se diferenciá-los pelos objetivos da análise realizada. Enquanto que em *data mining* o objetivo é identificar padrões ainda não conhecidos entre variáveis imputadas, em *machine learning* o objetivo é aprender continuamente com novos dados com vistas a melhorar a previsão sobre uma determinada variável. Assim, em *machine learning* o pesquisador irá incluir, dentre os atributos para gerar o modelo, variáveis que já são suspeitas de causar impacto na variável alvo e o objetivo é, dado esses atributos, avaliar qual o *output* que o modelo retorna para cada instância da base de dados (vetor com os atributos selecionados). Num trabalho de *data mining* atributos diversos sobre os quais suas relações são desconhecidas são avaliados, em busca de relações ocultas, previamente desconhecidas. Portanto, considerando o objetivo desse trabalho, pode-se dizer que trata-se de uma aplicação de *machine learning.*

2.6. Machine learning como ferramenta de CRM

A partir da década 2000, o avanço da computação começa a viabilizar aplicações práticas de *machine learning* em ambientes de negócio. Notoriamente o Google tem, como diferencial competitivo que alavancou seu sucesso, um sistema de busca inteligente, capaz de aprender o que o usuário procura com base em experiências de uso passadas, exibindo resultados cada vez mais relevantes (ROUSSEAU, 2010). Quando combinado a um modelo de negócios de anúncios

[1] A tradução mais comum em português para o termo *machine learning é* aprendizado de máquinas. Todavia, este estudo utiliza a nomenclatura em inglês, por considerar ser mais usual, mesmo em textos em português.

pagos, direcionados a usuários também com base em modelos associativos de *machine learning*, o Google revolucionou o mercado de publicidade online, tornando-se o que é hoje. Em outro exemplo do uso dessa tecnologia como forma de proporcionar uma melhor experiência para o consumidor, a Amazon, maior varejo online dos EUA, utiliza dados gerados em seu próprio ambiente para, dentre outros serviços, sugerir produtos que possam interessar a usuários do site, de maneira individualizada. Assim, cada cliente recebe recomendações relevantes com base em seu próprio perfil, que foi identificado por um algoritmo com base em compras passadas desse usuário e de outros com gostos afins – e cuja afinidade também foi identificada por uma máquina (AL IMRAN, 2014). Essas empresas, na vanguarda tecnológica em seus mercados, souberam aproveitar o poder do aprendizado de máquinas para aplicações em seus negócios, transformando-o em diferencial competitivo para oferecer experiências melhores para seus consumidores.

Mas, diferente do que esses exemplos possam sugerir, a aplicação de *machine learning* para melhorar a interação com os consumidores não é uma exclusividade de empresas grandes e complexas. Negócios de portes variados podem aplicar modelos para ajudar em cada uma das quatro atividades básicas do CRM: identificar, diferenciar, interagir e customizar. Particularmente, este trabalho foca em sua aplicação para diferenciação de clientes (por nível de risco de evasão) e a academia mostra que é possível utilizar dados naturalmente gerados nos ambientes de negócio para extrair informações relevantes para esse fim.

Uma indústria com particular interesse em desenvolver modelos preditivos para identificação precoce de clientes com possíveis comportamentos desviantes no futuro é a bancária e por isso, aparece como *early adpoter*, com casos de aplicações nessas indústrias descritos na academia ainda no início da década de 2000. Uma aplicação óbvia é a previsão de risco de crédito dos clientes aos quais o banco está exposto. West (2000) utiliza bases de dados de *score* de crédito de clientes de instituições australianas e alemãs para sugerir melhorias na previsão da classificação de risco de default por meio de modelos de redes neurais. Ainda uma novidade naquele ano, o autor discute problemas relacionados aos modelos estatísticos paramétricos tradicionais e sugere que a implantação de um modelo de

machine learning, não paramétrico, poderia melhorar, ainda que marginalmente, a qualidade das previsões. Zekikc-Susac et al. (2004) utilizam dados de um banco croata para linhas de crédito a pequenas empresas para classificar preventivamente se o cliente que busca um empréstimo seria um bom ou mal pagador - definido como aquele atrasa por mais de 45 dias uma parcela da dívida. As autoras testam diferentes metodologias de *machine learning* e chegam a 81% de acurácia com um algoritmo de redes neurais.

2.7. Aplicações de machine learning no setor de educação

Especificamente no caso do setor de educação, técnicas de *machine learning* encontram no problema da evasão de alunos uma oportunidade eficaz de aplicação. Os dados gerados usualmente sobre alunos, relacionados a desempenho acadêmico atual e pregresso, informações demográficas coletadas no ato da matrícula e informações do relacionamento financeiro com a instituição podem indicar padrões associados ao abandono da instituição. De fato, há casos publicados na academia internacional e brasileira que demonstram o potencial desta abordagem como instrumento de identificação de alunos em risco, para que a gestão das instituições de ensino possam atuar preventivamente. A seguir, são resumidos quatro trabalhos acadêmicos recentes, dois americanos e dois brasileiros, onde são aplicadas metodologias similares.

2.7.1. Estudo 1: Aulck et al (2016) – Universidade de Washington

A alta taxa de evasão no ensino superior não é uma problemática exclusiva do Brasil. Em Aulck et al (2016), os autores apontam que a taxa de evasão no primeiro ano nos EUA é de cerca de 30%, comparável, portanto, à realidade brasileira. Nesse estudo, os autores utilizam métodos de *machine learning* para prever o risco de evasão de alunos a partir de uma base de 32,5 mil alunos de uma grande instituição pública, a Universidade de Washington – que os autores apontam como a maior base de dados conhecida já utilizada para este fim. Os atributos utilizados foram variáveis demográficas (idade, gênero, raça/etnia, status de residência), notas de desempenho acadêmico anteriores ao ingresso na universidade (provas SAT e ACT) e dados acadêmicos dos alunos no primeiro ano da graduação

(notas, disciplinas cursadas, em que momento foram cursadas). Não foram utilizadas informações financeiras, como custo dos cursos, se o aluno é bolsista ou utiliza algum tipo de financiamento – o que foi apontado pelos autores como uma das limitações do estudo.

A partir dos dados acadêmicos básicos, uma série de novos atributos foram criados, como a contagem do número de disciplinas cursadas de um determinado departamento, o CR acumulado apenas de matérias de um mesmo departamento, assim como a contagem de disciplinas cursadas e CR acumulado em disciplinas tidas mais difíceis – os autores selecionaram determinadas disciplinas em áreas de ciência, tecnologia, engenharia e matemática consideradas *gatekeepers*, ou seja, matérias com notória dificuldade entre os estudantes. Importante ressaltar que essas disciplinas não foram identificadas estatisticamente, e sim, com base em conhecimento prévio de características dos cursos pelos autores.

Três tipos de algoritimos foram utilizados e comparados, tendo índices de acurácia de previsão próximos: regressão logística (66,59%), *random forests* (62,24%) e *k-nearest neighbors* (64,60%). Embora individualmente nenhum dos atributos tenha tido uma acurácia de previsão superior a 54%, dois tipos de atributos se sobressaíram como bons indicadores: CR em disciplinas de determinados cursos (matemática, inglês, química e psicologia) e a data de nascimento e semestre de matrícula na instituição. O segundo grupo é explicado por tendência de queda ao longo do tempo na taxa de evasão da instituição: a evasão de alunos no primeiro ano em 1998 era de 27,6%, contra 20,2% em 2006, por exemplo. Embora os autores não busquem explicar por que o CR de determinados cursos é mais relevante do que de outros para prever a evasão, a conclusão do estudo é que mesmo com a análise de desempenho acadêmico de um único período letivo, é possível gerar um modelo de previsão com acurácia relativamente alta a partir de algoritmos de *machine learning*.

2.7.2. Estudo 2: Delen (2010) – IES Pública nos EUA

Em Delen (2010), o estudo tem como objetivo criar um modelo para identificar estudantes do primeiro período da graduação com alto risco de evasão e identificar as variáveis mais relevantes para esse tipo de previsão, para ser utilizado

como ferramenta de apoio à gestão da instituição de ensino. A instituição alvo de seu estudo foi uma universidade pública do Meio Oeste americano, com cerca de 23 mil alunos matriculados. A variável escolhida como classe foi binária ("retornou para o segundo período/não retornou".

O autor coleta variáveis relevantes para o problema tendo como base a teoria da integração do aluno de Tinto (1975, 1987, 1993), que mostra que o desempenho acadêmico atual e histórico tem grande poder de previsão sobre o desempenho futuro e sobre a capacidade de persistência dos alunos. Assim, Delen incorpora no modelo a média final escolar, a nota no SAT e o CR acumulado no primeiro ano da graduação. Ainda tomando Tinto como referência, o autor busca como atributos variáveis que captem a integração do aluno com a instituição. Por isso inclui se o aluno possui algum *major*[2] e se mora no mesmo estado da IES – já que alunos que moram longe podem se sentir desintegrados ou emocionalmente desconectados da comunidade onde estudam. Além dessas, o autor também inclui no estudo variáveis financeiras, por apontar outros estudos que indicam correlação entre bolsas e nível de comprometimento com estudos e, por outro lado, dívidas de financiamento estudantil com baixo rendimento acadêmico.

Além dessas, outras variáveis acadêmicas, financeiras e demográficas foram utilizadas e, após tratamento dos dados, foram testados quatro algoritmos *machine learning*, regressão logística, redes neurais, árvore de decisão e *support vector machines*. Os resultados apresentados mostram que é possível ter alto índice de acerto na previsão da evasão. Os modelos gerados por todos os métodos apresentaram acurácia elevada, acima de 70%, tendo o último 81% de acurácia na previsão da classe. Portanto, o autor evidencia a eficácia da aplicação de uma metodologia de processamento de *big data* para produzir informação gerencial relevante para a retenção de alunos, a partir de dados que já são naturalmente produzidos pelas instituições ou aos quais elas já têm acesso.

[2] Nos EUA, os alunos podem escolher áreas de especialização, chamadas de *majors*, dentro dos seus cursos de graduação. Para obter um Major é preciso cumprir um determinado número de disciplinas dentro da área escolhida. Como não são obrigatórios, a intensão de Delen foi tomar essa variável como uma *proxy* para o engajamento do aluno com seu curso e sua formação.

2.7.3.
Estudo 3: Balaniuk et al (2011) – IES privada premium em Brasília

A academia no Brasil possui casos de aplicações de modelos de *machine learning* para prever evasão de alunos no ensino superior no país. Balaniuk et al (2011) utilizam dados internos de 11.495 alunos de uma IES privada *premium* de Brasília para chegar num modelo de previsão de evasão. Os autores utilizam algoritmos de árvore de decisões, regressão logística e redes neurais que depois são combinados num único modelo. As variáveis observadas pelos autores foram (i) grupo de idade, (ii) gênero, (iii) bairro de residência, (iv) status de trabalho, (v) tipo de escola de ensino médio frenquentada, (vi) renda familiar, (vii) CR médio, (viii) CR do segundo período, (iv) presença média total, (x) presença média no segundo período e (xi) número de reprovações nos dois primeiros períodos.

O bairro de residência foi tomado como *proxy* para a situação de renda familiar, complementando a variável (vi). Grupo de idade, gênero e status de trabalho foram incluídas para tentar medir possíveis incompatibilidades de horário – os autores esperavam que pessoas mais jovens e que não trabalham tivessem menor propensão a ter problemas de horário para estudar, assim como mulheres com crianças ou grávidas poderia ter um problema de longo prazo para se comprometer a um determinado horário. Já a escolha da variável tipo de escola, dividida entre pública, privada ou militar, buscou avaliar possíveis deficiências na educação básica do aluno. Por fim, as variáveis de CR, presença e reprovações foram escolhidas como medidas para o sucesso acadêmico. Os autores ponderaram as médias pelos pesos dos créditos das disciplinas e, para as reprovações nos dois primeiros períodos, criaram uma variável de três valores: nenhuma reprovação, até duas reprovações e três ou mais.

A amostra de alunos foi dividida em dois grupos: 3.058 que concluíram a graduação ou evadiram e 8.437 alunos que não se formariam ao fim do último semestre disponível na base de dados, estavam ao menos no terceiro período e tinham completado uma pesquisa de dados socioeconômicos da IES. Enquanto o primeiro grupo foi utilizado como base para treino (2/3) e testes do modelo (1/3), o segundo foi usado para gerar as previsões, sendo um modelo de classificação binário, com dois possíveis outputs: conclusão ou evasão, juntamente com as

probabilidades estimadas para cada aluno classificado. A acurácia total do modelo, dada a partir dos testes, indicou 80,6% de índice de acerto na classificação. O estudo aponta que numa análise preliminar, coordenadores e professores da IES concordaram com alguns alunos identificados como de alto risco de evasão por conhecerem seus casos e que, de fato, na renovação de matrícula para o semestre posterior ao estudo, muitas evasões previstas ocorreram – embora não tenham detalhado mais informações sobre a predição contra as observações.

O estudo conclui que com essa ferramenta, que indicou que 38,5% da base de alunos do grupo 2 iria evadir, a IES tem à mão uma lista seleta daqueles com maior risco de abandonar a instituição. Com essa informação na mão, ela pode escolher em quem focar seus esforços de retenção. Alunos com probabilidade muito alta de evasão permitem duas possíveis decisões do ponto de vista de CRM: ou a IES não faz nada, pois os esforços serão em vão, ou tenta reverter a situação, analisando individualmente cada caso e oferendo soluções compatíveis com o problema que levaria o aluno abandonar o curso, como suprir uma deficiência acadêmica, orientação vocacional ou mesmo auxílio financeiro. Seja qual for a linha de ação, os autores reconhecem que a melhor forma de prevenir a evasão é no contato direto e pessoal com os alunos e que, com uma relação mais próxima com a IES, buscando entender e suprir as necessidades dos alunos, desenvolve-se lealdade. Nesse sentido, a metodologia aplicada pelo estudo teve como objetivo apresentar apenas uma ferramenta para auxiliar a instituição a escolher quais alunos devem receber esse tipo de atenção.

2.7.4. Estudo 4: Silva & Adeodato (2012) – UFPE, IES pública federal de Pernambuco

Silva & Adeodato (2012) utilizam dados da Universidade Federal de Pernambuco (UFPE) para construir um modelo de *machine learning* com propósito similar de previsão de comportamento futuro dos alunos. No entanto, por se tratar de uma IES federal, o estudo analisa outro fenômeno tido como problema na educação pública, além da evasão: o que chamam de retenção, definido pelos autores como o aluno que demora mais tempo para se formar do que o programado em sua estrutura curricular. Como não há contrapartida de receita incremental por cada aluno matriculado na IES pública – embora haja custos associados a cada

aluno – um maior tempo de formação implica em maior custo por formando. Além disso o atraso de alunos reduz o número de vagas disponíveis nas IES públicas. O impacto econômico do maior tempo para graduação é mitigado na IES privada pois o aluno continua a pagar regularmente por seus estudos. Portanto, considerando um problema específico da IES pública para o qual sua gestão tem metas de redução, os pesquisadores, eles próprios professores da instituição, desenvolveram no estudo um modelo para prever a probabilidade de retenção de alunos após a conclusão do segundo período, para que a IES possa agir preventivamente ao longo do curso, garantindo a formação do estudante no tempo previsto em sua grade curricular.

Assim como em Balaniuk et al (2011), Silva e Adeodato definiram como corte na amostra analisar apenas estudantes que concluíram os dois primeiros semestres de seus cursos de graduação – pois utilizam informações produzidas nesses dois períodos como atributos – e analisaram apenas aqueles que iniciaram e concluíram sua atividade acadêmica na UFPE entre 1998 e 2008 em 6 cursos selecionados. Removendo *outliers* chegaram a uma amostra de 5.793 aluno, dos quais 2.723 se enquadravam como alunos retidos. As variáveis primárias foram coletadas a partir de uma estrutura de dados por curso/semestre/disciplina/aluno. Os autores transformaram para uma base de dados com as matrículas como instância, contendo as informações de curso, ano e semestre de entrada, disciplinas cursadas, desvio da disciplina em relação ao semestre em que deveria ser cursada, status final do aluno na disciplina e notas finais na disciplina.

A partir dessas variáveis criaram outras que foram incluídas como atributos a serem analisados pelo modelo: média do CR no primeiro e segundo semestres; taxas de presença e de reprovações nas disciplinas dos dois primeiros períodos, além de o número absoluto de reprovações; variação da taxa de reprovações do primeiro para o segundo período; taxa de aprovação na prova final de cada semestre, número de trancamentos de disciplinas no segundo período; número de disciplinas matriculadas no segundo período (no primeiro, todas são obrigatórias); coeficiente de variação das notas nos dois primeiros semestres e; duas variáveis binárias indicando se o aluno reprovou nas disciplinas mais importantes do primeiro e do segundo períodos – essa variável foi calculada a partir da identificação da disciplina

que mais reprovou alunos, para cada curso em cada semestre, ou seja, considerando como *proxy* para a importância a dificuldade de aprovação na disciplina.

Os autores optaram por rodar um modelo de regressão logística por considerarem que este tipo algoritmo funciona bem para classificação de variáveis binárias dependentes de outras variáveis independentes entre si. Além disso, utilizaram algoritmos de indução de regras (LEE; STOLFO; MOK, 1999) para chegar a atributos independentes que sozinhos ou em conjunto com outros, possuem significância estatística em explicar uma alteração na classe investigada. A relevância das regras induzidas foi testada pela métrica conhecida como *lift* – a razão entre a frequência relativa da classe-alvo com a regra e a frequência na amostra como um todo. Algumas regras interessantes surgem como estatisticamente significantes para determinar a probabilidade de retenção de um aluno na IES, tal como definido pelos autores. A regra associativa mais relevante para prever a classe foi "reprovar na matéria mais importante do segundo período" com "taxa de trancamento de disciplinas no segundo período entre 25% e 50%". Outras regras que surgiram com *lift* alto foram "CR do segundo semestre < 3"; e "curso = economia" com "número de disciplinas cursadas no segundo período < 3".

Ao fim do estudo, os autores calcularam o custo total da UFPE por aluno por semestre (R$1.672) e compararam com o custo de um programa de aconselhamento estudantil para estudantes a partir do terceiro período. O objetivo foi avaliar a viabilidade econômica de sua implantação, comparando o custo de manutenção de tal programa com diferentes cenários de economias de custos associados a reduções na taxa de retenção. A conclusão foi que, caso o programa tenha sucesso em reduzir a taxa de retenção em ao menos 15%, a decisão de implantação é economicamente viável. Os autores encerram o estudo destacando, dentre outras, a necessidade de incorporação de variáveis sócio demográficas e de inputs comportamentais pelos próprios professores em modelos futuros.

Como pode ser observado pelos estudos apresentados anteriormente, há interesse crescente da academia em aplicações práticas de *machine learning* preditivo para o setor de educação superior, pautados em dados já produzidos neste ambiente. Todavia, as diferenças nos tamanhos amostrais entre os dois estudos realizados nos Estados Unidos – Aulck et al (2016) com 32 mil alunos e Delen

(2010) com 23 mil alunos – vis à vis os estudos realizados no Brasil – Balaniuk et al (2011) com 3 mil alunos e Silva e Adeodato (2012) com 2,7 mil alunos – evidenciam escalas de aplicação diferentes. As datas de publicação dos estudos também mostram que este tipo de abordagem para o problema da evasão é relativamente recente, sendo um campo promissor que apresenta muitas possibilidades de investigação a serem exploradas.

3.
Metodologia

Este estudo aborda a evasão estudantil no ensino superior privado como um problema de gestão de marketing de relacionamento. Para tanto, é necessário avaliar informações específicas de clientes, permitindo que estes sejam tratados de forma individualizada, o que aumenta a eficácia das ações de marketing, particularmente ações de retenção. Optou-se neste estudo pela utilização de uma abordagem quantitativa, usufruindo-se do acesso aos conjuntos de dados de sistemas financeiro e acadêmico de uma IES, disponibilizados com exclusividade por sua equipe de gestores especificamente para este trabalho. Bases de dados internas de clientes de empresas podem proporcionar ricas oportunidades de análise de comportamento do consumidor, porém em geral a academia possui escasso acesso a esse tipo de informação para a atividade de pesquisa.

A seguir, são apresentados o tipo de pesquisa, os dados coletados, os procedimentos utilizados para tratar e analisar os dados e o processo da amostragem.

3.1.
Tipo de Pesquisa

Este estudo é dividido em três etapas. Na primeira, utilizou-se uma abordagem quantitativa de análise, aplicando modelos de *machine learning* de classificação a uma base de dados secundários para encontrar modelos que expliquem a evasão de alunos. Na segunda, aplica-se os modelos encontrados a uma amostra relevante de alunos ativos, que são, então, classificados de acordo com sua probabilidade de evasão. Por fim, aplicou-se o conceito de *lifetime value* a essa base de alunos, com uma estimativa de valor econômico para a instituição de cada aluno e da base como um todo. Os dados refletem uma amostra representativa do universo de alunos da IES que ingressaram a partir de 2010.

Para a primeira e segunda etapas, a amostra foi dividida em três grupos: (i) alunos que concluíram seu curso, (ii) alunos que evadiram e (iii) alunos que ainda estavam ativos no momento da coleta dos dados. Os dados foram dispostos como uma matriz de informações onde cada linha, chamada de instância, representa uma

única matrícula e cada coluna representa uma variável diferente relacionada àquela matrícula. Para rodar modelos de classificação, requer-se que uma das variáveis seja denominada como a "classe". Classe é a variável-alvo para a qual deseja-se aprender sobre, para que futuros dados, ainda sem classe definida – tal como o grupo de alunos (iii) – possam ser classificados pela máquina como pertencentes a uma das possíveis classes. Como o objetivo deste estudo é gerar um modelo preditivo de classificação do risco de evasão para alunos ainda ativos definiu-se como classe a variável binária "evade/conclui". Portanto, esta variável assume o valor "conclui" para os alunos do grupo (i) e "evade" para os alunos do grupo (ii). As outras variáveis disponíveis para cada instância são chamadas de atributos, e representam as características de cada matrícula que foram avaliadas pelos algoritmos.

Os grupos (i) e (ii) compreendem a amostra para a qual os modelos de *machine learning* foram calculados na primeira etapa, enquanto o grupo (iii) foi isolado pois são dados não classificados. Esse grupo foi usado na segunda etapa da pesquisa. Utilizando o WEKA (HALL, 2009; AHER; LOBO, 2011; SHARMA; JAIN, 2013), um *software* específico para desenvolver modelos de aprendizado de máquinas, aplicou-se cinco tipos de algoritmos de classificação, que serão indicados posteriormente. De forma genérica, o que esses algoritmos fazem é avaliar os atributos de cada instância e sua respectiva classe para desenvolver um modelo que as explique. Uma vez desenvolvidos os modelos na primeira etapa, estes foram aplicados ao grupo (iii), de alunos ativos, para serem classificados a partir de seus atributos, considerando o que foi aprendido com a amostra composta dos grupos (i) e (ii). Por fim, na terceira etapa calculou-se o *lifetime value* de cada aluno ativo ponderando-os de acordo com as probabilidades geradas pelo modelo na segunda etapa. Uma consequência importante dessa metodologia é que os resultados dos modelos não são capazes de dizer em que momento o aluno irá evadir. Os processos de classificação apenas indicarão a probabilidade de evasão do aluno em algum momento, dadas as informações produzidas nos dois primeiros períodos.

3.2.
Dados Coletados

Na tabela 3.1 a seguir estão dispostas as variáveis coletadas nos bancos de dados da IES. Esses foram os dados obtidos primariamente.

Tabela 3.1. – Tabela de Dados Coletados.

Categoria	Informação	Comentário
Sócio Demográficas	Nome completo	
	Número de Matrícula	Todas as informações dos banco de dados acadêmicos estão associadas pelo número de matrícula, por isso este foi utilizado como código identificador de cada aluno neste estudo.
	Data de Nascimento	Dado utilizado para calcular a idade no ano de ingresso
	Gênero	Masculino ou feminino
	Endereço de Domicílio	Das informações de endereço, aproveitou-se apenas o CEP, que foi utilizado para medir a distância até a IES
Acadêmicas	Notas das provas	Notas das avaliações das disciplinas cursadas, de zero a 10.
	Nome e código do curso	Foram analisados alunos de 20 cursos. Detalhe na lista seguinte.
	Tipo de Ingresso	Os tipos possíveis de ingresso em cursos da IES são: (i) transferência externa ou (ii) interna, (iii) nota no ENEM, (iv) vestibular interno, (v) via PROUNI ou (vi) por já ser diplomado em outro curso da IES.
	Data de Ingresso	Ano e semestre em que cursou o primeiro período no curso.
Financeiras	Títulos emitidos	Histórico de todos os títulos de cobrança emitidos contra alunos da instituição, com informações de data de vencimento, pagamento, valor de face e valor pago, desconto em relação ao "preço cheio" do curso dentre outras. Em geral, equivalem à mensalidade, mas há exceções, como alunos que optam por pagamento semestral ou títulos de renegociação de dívidas com a IES.
	Aluno FIES	Aluno possui FIES: não/sim

Foram considerados para o estudo alunos de cursos de bacharelado, licenciatura e tecnólogos, com durações que variam de 4 a 10 semestres. Abaixo segue a lista dos cursos analisados com os respectivos períodos de duração e tamanho da amostra em cada caso.

Tabela 3.2. – Tamanho da amostra de alunos por curso e duração

Curso	Sigla	Duração (sem.)	Amostra
Psicologia	PS	10	444
Enfermagem	EN	10	686
Educacao Física	EF	8	864
Biologia	BI	8	267
Administração	AA	8	397
Fisioterapia	FS	10	91
Tecnólogo em Gestão de Recursos Humanos	TR	8	354
Ciências Contábeis	AC	4	183
Engenharia Ambiental e Sanitária	EA	8	124
Nutrição	NU	10	132
Farmácia	FM	8	337
Fonoaudiologia	FO	8	25
Tecnólogo em Estética e Cosmética	TE	10	97
Tecnólogo em Gestão de Marketing	TM	4	33
Tecnólogo em Processos Gerenciais	TG	4	6
Total			4040

3.3. Procedimentos de Tratamento e Variáveis Utilizadas

Algumas informações extraídas dos bancos de dados da instituição foram utilizadas diretamente na base de dados final, para a qual os modelos foram calculados. Outras variáveis foram calculadas a partir dos dados existentes. A lista a seguir contém todos os atributos presentes na base final, e suas respectivas descrições:

Tabela 3.3. – Tabela de variáveis utilizadas

Nome da Variável	Categoria	Tipo de dado	Descrição
matricula		real	Número identificador do aluno
ingresso	Acadêmica	string	Tipo de Ingresso: TRANSFERENCIA_EXTERNA, ENEM, PORTADOR_DE_DIPLOMA, VESTIBULAR, PROUNI, TRANSFERENCIA_INTERNA
fies	Financeira	real	Aluno possui FIES: não/sim
distancia	Sócio Demográfica	real	Distância em Km do endereço cadastrado para a IES - calculado a partir do CEP via API do Google Maps
curso	Acadêmica	string	Código identificador do curso. Ao todo, a base possuía 20 cursos únicos, incluindo alguns que não são mais ofertados pela IES
gênero	Sócio Demográfica	string	M/F
turno	Acadêmica	string	Manhã/Tarde/Noite
tipo_colegio	Demográfica	string	Tipo de escola em que estudou o último ano: Público/Privado
anos_escola	Demográfica	real	Número de anos entre a entrada na IES e a formatura do Ensino Médio
idade	Demográfica	real	Idade no ano de ingresso na IES
media_semestre_1	Acadêmica	real	Média das disciplinas no 1º semestre cursado
media_semestre_2	Acadêmica	real	Média das disciplinas no 2º semestre cursado
melhorou	Acadêmica	real	1 se a média do 2º semestre foi maior que o do 1º e 0 caso contrário
desconto_medio_1	Financeira	real	Desconto médio dos boletos no 1º semestre
desconto_medio_2	Financeira	real	Desconto médio dos boletos no 2º semestre
reduziu_desconto	Financeira	real	1 se o desconto do 2º semestre foi menor que o do 1º e 0 caso contrário
atrasos	Financeira	real	Percentual de boletos pagos em atraso/total de boletos emitidos
negociacoes	Financeira	real	Percentual de boletos renegociados/total de boletos emitidos
isencoes	Financeira	real	Percentual de boletos isentados / total de boletos emitidos - se o aluno é bolsista integral, sempre terá isenções. Eventualmente isenções são dadas por outras razões, como isentar o boleto de janeiro para um aluno que só se matriculou em março. Nesse caso, o boleto de janeiro é emitido, mas isentado.
status	Classe	classe	Esta é a variável que define o status do aluno: evade/conclui

3.4. Processo de amostragem

Dois tipos de amostra foram necessários para a realização deste estudo. A primeira consiste em alunos que concluíram sua graduação na IES ou que evadiram. Essa amostra, portanto, compreende dados cuja classe é conhecida, (evadiu ou concluiu). A segunda é a amostra de alunos ainda ativos na instituição, cuja classe ainda é, por definição, desconhecida.

A data inicial de corte escolhida foi o primeiro semestre de 2010 devido a limitações de acesso à dados históricos anteriores a esse período no sistema da IES. Algumas premissas foram assumidas no momento da classificação do aluno como concluinte ou evadido, de acordo com os tipos de status encontrados no sistema acadêmico da IES e são descritos a seguir.

3.4.1. Status e Critérios de Classificação

O primeiro passo de separação dos alunos exigiu a identificação dos seus status de acordo com o sistema acadêmico. Toda a extração de dados para esse trabalho ocorreu em agosto de 2016, portanto durante a janela de matrícula de alunos para o segundo semestre desse ano. Assim alguns dos status exibidos são temporários, como o status “aguardando renovação” – nesse caso em particular, significa que o aluno concluiu o semestre anterior, ainda possui matérias a cursar e está apto a se matricular para o semestre atual, mas ainda não o fez. Após o fechamento da janela de matrícula, caso o aluno não a tenha renovado, seu status é automaticamente atualizado para trancado. Caso tenha renovado a matrícula no período, seu status passa para cursando. Neste caso em particular, não é possível saber se o aluno evadiu ou não, no entanto sabe-se que ele estava ativo em 2016.1. Portanto, este trabalho considera como o último período disponível 2016.1 e, para o caso citado, o aluno é classificado como ativo.

A tabela 3.4 indica todos os status associados aos alunos, encontrados no sistema, e as respectivas classificações adotadas para este estudo.

Tabela 3.4. – Tabela de status acadêmico e classificação equivalente.

Status Sistema	**Classificação**
CURSANDO	Ativo
AGUARDANDO RENOVAÇÃO	Ativo
MATRÍCULA ACADÊMICA	Ativo
NÃO FORMOU TURMA	Ativo
COLOU GRAU	Conclui
CONCLUINTE	Conclui
DEVE ATIVIDADE COMPL.	Conclui
DEVE ENADE	Conclui
MATRÍCULA IRREGULAR	Desconsiderar
FALECIDO	Desconsiderar
MATRÍCULA TRANCADA	Evade
TRANSFERÊNCIA INTERNA	Evade
TRANSFERÊNCIA	Evade
MATRÍCULA CANCELADA	Evade
TRANSFERÊNCIA EXTERNA	Evade
MATRÍCULA TRANCADA/NÃO RECOMPRA	Evade
MATRÍCULA CANCELADA/NÃO RECOMPRA	Evade
CANCELADO	Evade
TRANCADO	Evade

"Matrícula acadêmica" é um status transitório, similar a "aguardando renovação". A diferença reside no fato de que neste caso o aluno já fez a chamada matrícula acadêmica mas ainda não pagou o boleto referente ao primeiro mês do semestre. Pelo critério da instituição, o aluno só oficializa sua matrícula ao realizar a chamada matrícula financeira. Esta ocorre com o pagamento do boleto de mensalidade referente ao mês de janeiro ou de julho, para o primeiro e segundo semestres, respectivamente. Portanto, no caso desse status, o aluno já sinalizou que

pretende continuar na instituição, teve o boleto emitido, mas este ainda não foi computado como pago.

O status “não formou turma” também é transitório e refere-se a alunos que apesar de estarem matriculados, ainda não estão alocados a nenhuma turma, pois as matérias em que se inscreveu foram canceladas e ele não estava associado a nenhuma no momento da extração dos dados do sistema. Em todos esses casos, o aluno estava ativo em 2016.1, que é o último período considerado na análise deste trabalho. Por isso, foram classificados como ativos para o fim de separação das amostras.

Para os alunos classificados como “conclui”, este estudo considerou, além daqueles alunos que terminaram e colaram grau ou não, aqueles que ainda devem atividade complementar ou realizar a prova do ENADE, necessários para conclusão do curso. Este estudo assume a premissa de que, nesses casos, apesar de o aluno ainda não ter oficialmente terminado seu curso, ele está em vias de termina-lo, pois já passou por todo o ciclo acadêmico, cursando todas as disciplinas necessárias.

O status matrícula irregular foi encontrado em alguns casos e descartados, pois verificou-se que trata-se de problemas cadastrais no sistema. Além desse caso, alunos que faleceram ao longo do curso também foram excluídos deste estudo. Todos os demais status representam alunos que evadiram, por uma razão ou outra, como por transferência interna de curso, por trancamento espontâneo, ou por simples abandono.

3.4.2. Critérios de amostragem para alunos com classe conhecida

Uma vez identificados os alunos de acordo com os critérios estabelecidos neste estudo, optou-se por um filtro para obter apenas aqueles alunos que cursaram ao menos dois períodos na IES antes de evadir, seguindo assim critério similar ao utilizado por Balaniuk et al (2011) e Silva e Adeodato (2012). Essa escolha foi tomada por razão similar: com apenas um semestre cursado, há poucas informações produzidas sobre o aluno, o que enfraquece a capacidade de aprendizagem e predição de algoritmos de *machine learning.* É importante ressaltar que a evasão ocorre principalmente no primeiro ano, no entanto por conta da metodologia

utilizada neste estudo, seus resultados serão úteis como ferramenta de gestão para retenção de alunos que cursaram ao menos dois períodos, tal como nos estudos supracitados.

O sistema acadêmico utilizado pela instituição possui a limitação de só exibir o status atual do aluno, mas não seu histórico de mudanças. Portanto, caso um aluno tenha trancado sua matrícula, mas depois reaberto, esse trancamento não está registrado no sistema. Porém, indiretamente é possível saber se o aluno chegou a se matricular para um determinado semestre com base em informações do sistema financeiro. Como a formalização da matrícula ocorre com o pagamento da mensalidade de janeiro ou de julho (de acordo com o semestre), é possível cruzar esses dados para checar se um aluno chegou a estar ativo em determinado semestre. Neste estudo, foram considerados apenas alunos que cursaram os dois primeiros períodos de seu curso de forma consecutiva, processo que foi validado de acordo com uma variável binária criada a partir da base financeira assumindo o valor de 1 se as mensalidades de janeiro/X e julho/X, ou julho/X e janeiro/X+1 foram pagas – onde X representa o ano de ingresso e a variável é calculada de acordo com o semestre inicial (primeiro semestre do ano, ou o segundo, respectivamente) – e zero caso contrário. Alunos que não atenderam a esse critério foram excluídos da amostra. Portanto, no critério aqui utilizado, foram incluídos os alunos que cursaram os dois primeiros períodos, mas que não necessariamente concluíram o segundo, podendo a evasão ter ocorrido ao longo desse período. Ao final, chegou-se a uma amostra de 4.078 alunos classificados.

3.4.3. Amostra com classe desconhecida (Scoring)

Criar e treinar modelos e fazer previsões são dois processos distintos. No primeiro, utiliza-se uma base de instâncias com classe conhecida para que os algoritmos calculem e modelem as relações entre os atributos e a classe – no caso de técnicas de aprendizado supervisionado, como neste estudo. No processo de previsão, também conhecido como *scoring*, o objetivo é, já com o modelo calculado, aplica-lo sobre uma base de instâncias com classe desconhecida mas que possua observações sobre os mesmos atributos utilizados na base que originou o modelo. O output do modelo é um *score* para cada instância.

Para a amostra de alunos ainda ativos, optou-se por utilizar apenas aqueles que ingressaram na instituição em 2015.1 e 2015.2. Portanto, alunos que em 2016.2 deveriam ingressar no terceiro ou quarto semestres da IES. Para essa amostra foi aplicada a variável binária de validação de matrícula nos dois primeiros períodos, assim como na amostra com classe conhecida. Essa amostra tem o tamanho de 1.502 alunos e, além de classificadas pelos modelos, foram estimadas ainda as probabilidades de evasão (quando possível), assim como seu *lifetime value.*

3.5. Aplicação de algoritmos de Machine Learning

Este estudo utilizou o software de acesso gratuito WEKA para criar os modelos preditivos. Esse software possui um banco de algoritmos mais usados em processos de aprendizado de máquinas e permite criar, testar e comparar a qualidade modelos (HALL, 2009). O tipo de problema que este estudo trabalha é de classificação: reconhecer um aluno como pertencente a um dos grupos pré-definidos, no caso, "evade" ou "concluiu". Esse tipo abordagem é conhecido como aprendizado supervisionado, em oposição a metodologias de aprendizado não supervisionados. No primeiro caso, um modelo é criado a partir de dados já classificados, onde o algoritmo apenas aprende a relacionar os atributos àquela classe previamente definida. No segundo caso, não há classe definida e o algoritmo cria grupos a partir de associações "descobertas" – esse é o tipo de abordagem utilizado para construção de clusters por exemplo, ou de regras associativas.

Esse estudo foca em dois algoritmos para classificação, regressão logística (WILSON; LORENTZ, 2015; PENG; LEE; INGERSOLL, 2002) e árvore de decisão, cujos detalhes dos resultados são explorados no capítulo 5. Além desses dois, outros quatro algoritmos comumente utilizados para classificação também foram rodados e seus resultados comparados: J48 – outro tipo de árvore de decisão, também conhecido como C4.5 (QUINLAN, 1996; KOTSIANTIS, 2007) – K-Nearest Neighbors (ALTMAN, 1992), Support Vector Machines (CORTES e VAPNIK; FRADKIN e MUCHNIK, 2006) e *Naive Bayes* (HAND; YU, 2001; RICH, 2001).

As acurácias *out-of-sample* dos modelos e suas capacidades de generalização foram testadas com o método de validação cruzada conhecido como

k-fold cross-validation (WITTEN, FRANK e HALL, 2011), com k = 10. Essa metodologia consiste em dividir a amostra em k grupos de igual tamanho, onde k-1 grupos são utilizados para estimar os parâmetros, e 1 grupo é usado para teste de estimativa de classe. Em seguida, troca-se o grupo de teste, usando os demais k-1 grupos para estimação. Esse procedimento é realizado k vezes e, ao fim, tem-se uma estatística da quantidade de previsões sobre a classe corretamente realizadas sobre o total de previsões. Essa estatística, um número entre 0 e 1, é utilizada como medida da acurácia de um modelo de classificação e, naturalmente, quanto maior seu valor, melhor é o poder preditivo do modelo em questão.

3.6. Cálculo do Lifetime Value

Uma vez estimados os modelos, esses foram alimentados com a base de alunos com classe desconhecida, tendo como *output* a classe estimada (evade ou conclui) e uma probabilidade associada a essa classificação.

Para chegar ao *lifetime value* do aluno, aplicou-se a seguinte equação:

$$LV_x = p_x . \sum_{i=1}^{n} \frac{M_x . mgC}{(1+r)^i}$$

LV representa o *lifetime value* do aluno x; p representa a probabilidade estimada de o aluno x concluir o curso; n representa o número de meses até a sua formatura, M representa a mensalidade média do aluno x; mgC representa a margem de contribuição média da IES e; r representa a taxa de desconto representativa do custo de oportunidade para a IES.

Para chegar a n, calculou-se o número de períodos necessários até a conclusão do curso multiplicado por 6, obtendo-se o número de meses remanescentes esperado do aluno na IES, sem considerar atrasos na formação. Considerou-se também a mensalidade média paga pelo aluno, utilizando como proxy a mensalidade média do último semestre disponível, 2016.1. A margem de contribuição (58%) e taxa de desconto (15%a.a. nominal) utilizadas foram informadas pela direção da própria instituição. Este estudo optou por utilizar a margem de contribuição como proxy para o custo variável de servir o aluno. Dentre

os custos variáveis, o principal é o custo docente. Uma possível deficiência desse critério é que o custo docente é variável em função direta do número de turmas, não do número de alunos. No entanto, o número de turmas é, por sua, vez, função direta do número de alunos.

As probabilidades p de conclusão foram extraídas dos modelos de *machine learning* adotados. Os modelos produzidos pelos algoritmos *k-nearest neighbors* e e *support vector machines*, por limitações de suas próprias metodologias, não são capazes de fornecer essas estimativas probabilísticas, tendo como output apenas as classes. Portanto, a probabilidade de evasão para esses modelos é 0% ou 100%. Por isso, o *lifetime value* foi calculado utilizando as probabilidades obtidas a partir dos outros quatro algoritmos apresentados acima.

Por fim, somando-se o *lifetime value* de cada aluno, chega-se ao *customer equity value* dessa base de alunos ativos. Como foram calculados quatro valores, baseados em modelos distintos, comparou-se os resultados obtidos.

3.7. Limitações do estudo

Como a base de dados utilizada é de apenas uma IES, não é possível avaliar as características da própria instituição que contribuíram para a evasão. Para esse tipo de análise, seria preciso bancos de dados de outras IES para que os fatores intrínsecos de cada uma pudessem ser estatisticamente comparados. Outra limitação da metodologia utilizada é a impossibilidade de se prever exatamente em que momento a evasão ocorrerá. Assim, ainda que um aluno seja identificado como tendo risco elevado de evasão, os modelos não fornecem informações de em quanto tempo ela deverá ocorrer.

4.
Resultados

Esta seção é dividida em duas partes. Primeiro serão demonstrados e comparados os modelos obtidos. Na segunda parte, são exploradas as projeções e *lifetime value* obtidos.

4.1.
Modelos Gerados

Nas tabelas e gráficos a seguir os algoritmos estão representados pelos nomes pelos quais são chamados no software WEKA, de acordo com a tabela 4.1:

Tabela 4.1. – Nomes dos algoritmos para referência

Algorítmo	Nome Weka
Regressão Logística	Logit
Árvore de Decisão	REPTree
Árvore de Decisão J48	J48
K-Nearest Neighbours	IBk
Support Vector Machine	SMO
Naïve Bayes	N.Bayes

4.1.1.
Acurácia dos Modelos

As acurácias dos modelos são calculadas a partir de suas matrizes de confusão (FAWCETT, 2006). A matriz de confusão compara as classes reais com as previsões realizadas sobre as instâncias da base classificada. No caso deste estudo, como trata-se de um classificador binário, obtém-se uma matriz 2x2, tal como no exemplo da tabela 4.2:

Tabela 4.2. – Exemplo de matriz de confusão

		Previsão		
		0	1	Total
Real	0	35	5	40
	1	10	50	60
	Total	45	55	n=100

Os números na tabela representam contagens da amostra. Esse exemplo possui uma amostra de tamanho 100, sendo 40 instâncias pertencentes a classe “0” (que será referida com negativo) e 60 à classe “1” (positivo). A matriz também mostra o resultado das previsões e, relacionando essas informações temos que: (i) das 40 instâncias negativas, 35 foram corretamente classificadas (verdadeiro negativo), mas 5 foram incorretamente classificadas como positivo (falso positivo), erro conhecido como tipo I; (ii) analogamente, das 60 instâncias positivas, tem-se 50 corretamente classificadas (verdadeiro positivo) e 10 erros (falso negativo), chamados de erro tipo II. A acurácia total de um modelo é dada pela razão entre as previsões verdadeiras (células com fundo verde) e o tamanho total da amostra. Nesse caso, (35+50)/100 = 85%.

Além da acurácia total do modelo, outras estatísticas podem ser inferidas a partir da análise da matriz de confusão. Abaixo segue um breve descritivo das medidas que serão utilizadas neste estudo para fins comparativos.

Sensibilidade: também conhecido como *recall* ou taxa de verdadeiro positivo, mede o percentual de previsões de positivos, quando a classe real é positivo. No exemplo acima, 50/60 = 83%.

Especificidade: também conhecido como taxa de verdadeiro negativo, mede o percentual de previsões de negativo quando a classe real é negativo; 35/40 = 87,5%.

Precisão: é a taxa de acertos quando a previsão é positivo; 50/55 = 90.1%

Na tabela 4.3 são dispostas as matrizes de confusão de cada modelo, obtidos pelo método de validação cruzada 10 vezes. Aqui, “evade” representa a classe positivo.

Tabela 4.3. – Matrizes de confusão dos modelos calculados.

		Previsão		
Logit		Conclui	Evade	Total
Real	Conclui	2.377	290	2.667
	Evade	644	729	1.373
	Total	3.021	1.019	4.040

		Previsão		
REPTree		Conclui	Evade	Total
Real	Conclui	2.384	283	2.667
	Evade	653	720	1.373
	Total	3.037	1.003	4.040

		Previsão		
J48		Conclui	Evade	Total
Real	Conclui	2.304	363	2.667
	Evade	597	776	1.373
	Total	2.901	1.139	4.040

		Previsão		
IBk		Conclui	Evade	Total
Real	Conclui	2.149	518	2.667
	Evade	675	698	1.373
	Total	2.824	1.216	4.040

		Previsão		
SMO		Conclui	Evade	Total
Real	Conclui	2.428	239	2.667
	Evade	704	669	1.373
	Total	3.132	908	4.040

		Previsão		
N.Bayes		Conclui	Evade	Total
Real	Conclui	2.454	213	2.667
	Evade	846	527	1.373
	Total	3.300	740	4.040

A tabela 4.4 resume a acurácia, sensibilidade, especificidade e previsão dos modelos calculados.

Tabela 4.4. – Resumo dos resultados de desempenho dos modelos.

	Logit	**REPTree**	**J48**	**IBk**	**SMO**	**N.Bayes**	**Média**
Acurácia	**76,88%**	**76,83%**	**76,24%**	**70,47%**	**76,66%**	**73,79%**	**75,14%**
Sensibilidade	53,10%	52,44%	56,52%	50,84%	48,73%	38,38%	50,00%
Especificidade	89,13%	89,39%	86,39%	80,58%	91,04%	92,01%	88,09%
Precisão	71,54%	71,78%	68,13%	57,40%	73,68%	71,22%	68,96%
Evade/total	25,22%	24,83%	28,19%	30,10%	22,48%	18,32%	24,86%

A regressão logística e a árvore de decisão obtiveram os mais altos índices de acurácias dentre os algoritmos testados. Os métodos *support vector machines* e J48 também tiveram desempenho similar. Notadamente o algoritmo *support vector machines* demonstrou melhor desempenho na capacidade de prever corretamente a classe de evasão, com precisão de 73,7%, critério que evidencia também o pior desempenho do modelo baseado no algoritmo *k-nearest neighbors* (IBk), com 57,4% de precisão. As taxas relativamente altas de especificidade mostram, de

modo geral, baixo risco de se cometer o erro tipo I (alarme falso), todavia o risco de se cometer um erro tipo II (não identificação) é alto. Em geral, pouco mais de 50% de sensibilidade, indicando que quase metade dos alunos que evadiram não foram corretamente classificados. Vê-se portanto, uma dificuldade de identificar o aluno em risco baseando-se apenas nos atributos utilizados neste estudo. Embora não com a mesma intensidade, os modelos gerados em Balaniuk et al (2011) também mostraram maior chance de se cometer erro tipo II do que o tipo I. Talvez essa dificuldade se deva ao fato de que, em geral há certos atributos que, quando presentes, são sinais fortes de evasão (como um CR muito baixo, por exemplo) porém, sua ausência não significa que o aluno irá concluir o curso. Mesmo sem um CR baixo, um aluno pode evadir por outras razões não capturadas pelos atributos utilizados no estudo.

A prevalência da classe "evade" na amostra – o número de alunos que evadiram sobre o total amostral – é de 34%. A título de comparação, a tabela 5.3 indica o número de instâncias classificadas como "evade" sobre a amostra total. Todos os modelos exibiram um percentual menor da classe do que sua verdadeira prevalência. O algoritmo *k-nearest neighbors* foi o que chegou mais perto em termos quantitativos, no entanto, considerando a sua baixa precisão, chegou a esse número mais alto classificando mais instâncias erradas.

4.1.2. Análise do modelo de regressão logística

O método da regressão logística, cuja origem data do início do século XIX (CRAMER, 2002), é hoje um dos mais populares para classificação binária (WILSON; LORENTZ, 2015). Parte de seu atrativo é sua capacidade de indicar como as chances de ocorrência de um evento binário – a classe – são afetadas por um dado atributo. As chances são chamadas de *odds ratio.* Como exemplo ilustrativo para leitura dos seus resultados, se de cada 10 alunos evadidos, 7 são homens, então a probabilidade evasão de um homem é de 0,7 e a probabilidade complementar, de conclusão, é de 0,3. Nesse caso a razão 0,7/0,3 = 2,33 é a *odds ratio*. Portanto, nesse exemplo, pode-se dizer que homens tem 2,33 vezes mais chances de evadirem do que mulheres.

Como trata-se de uma razão entre probabilidades e estas oscilam entre 0 e 1, os possíveis valores de uma *odds ratio* variam de 0 a infinito positivo. Uma *odds ratio* entre 0 e 1 indica que aquele atributo está associado a uma menor chance de ocorrência da variável dependente. Uma razão igual 1 significa que o atributo não tem influência sobre a classe – o mesmo pode ser interpretado para números muito próximos de 1. Já valores maiores do que 1 indicam que o atributo aumenta as chances de ocorrência da classe e, quanto maior o valor, maior a chance. A tabela 4.5 lista os atributos e suas respectivas *odds ratios.*

Tabela 4.5. – Atributos e *odds ratios* estimados pelo modelo de regressão logística

Atributo	***Odds Ratio***
ingresso=ENEM	3.21
ingresso=PORTADOR_DE_DIPLOMA	0.50
ingresso=PROUNI	1.83
ingresso=TRANSFERENCIA_EXTERNA	0.47
ingresso=TRANSFERENCIA_INTERNA	2.39
ingresso=VESTIBULAR	1.72
fies	1.20
distancia	1.00
sigla_curso=AA	1.04
sigla_curso=AC	1.53
sigla_curso=BI	0.98
sigla_curso=EA	1.91
sigla_curso=EF	0.69
sigla_curso=EN	1.38
sigla_curso=FM	0.92
sigla_curso=FO	36.50
sigla_curso=FS	2.31
sigla_curso=NU	1.78
sigla_curso=PS	2.32
sigla_curso=TE	0.40
sigla_curso=TG	2.27
sigla_curso=TM	0.16
sigla_curso=TR	0.21
sexo=M	1.17
turno=MANHA	1.17
turno=NOITE	1.00
turno=TARDE	0.42
tipo_colegio=PUBLICO	0.77
anos_escola	1.00
idade	0.99
media_semestre_1	1.18
media_semestre_2	0.56
melhorou	0.62
desconto_medio_1	3.97
desconto_medio_2	0.94
reduziu_desconto	1.59
atrasos	1.26
negociacoes	1.12
isencoes	1.57
negociacoes	1.12
isencoes	1.57

É preciso ter cuidado com a interpretação desses resultados, pois os valores não necessariamente têm significado real, sendo apenas úteis para explicar a amostra analisada. Mas a partir dos resultados dispostos, é possível inferir algumas informações, ainda que com algum nível de incerteza. Por exemplo, o aumento da média das notas no primeiro para o segundo período reduz as chances de evasão, o que pode indicar um maior nível de engajamento do aluno ou maior nível de integração acadêmica no o conceito de Tinto.

Por outro lado, a redução do desconto médio tem um efeito de magnitude semelhante mas em sentido contrário: a perda de um benefício econômico pode ter o efeito de desmotivar o aluno. Curiosamente o nível de desconto no segundo período é muito pouco relevante (*odds ratio* próxima de 1), mas o nível de desconto do primeiro período tem relação direta com a evasão: ou seja, há uma correlação positiva entre alunos que entram com descontos maiores e evadem. Ao indicar que maiores níveis de desconto na entrada aumentam as chances de evasão o modelo chama a atenção para um possível viés de seleção na captação de alunos. Eventualmente descontos mais agressivos são dados para aumentar a captação, porém nesse processo é possível que a instituição acabe por trazer alguns alunos que se matricularam apenas porque a mensalidade estava barata, sem um vínculo maior com o curso ou própria instituição.

Alunos do FIES também aparecem com *odds ratio* maior que 1, indicando maior chance de evasão, mas neste caso é possível esse valor que esteja relacionado com um viés amostral: o FIES tornou-se relevante na instituição a partir de 2013, portanto não há tantos alunos formados no período, mas há abandonos.

A forma de ingresso na IES também parece ter um papel importante: alunos transferidos de outras instituições tem menor chance de abandonarem o curso. Esse fato também pode estar relacionado à maior motivação do aluno com aquele curso em particular, pois este, apesar de ter tomado a decisão de mudar de instituição, por uma razão qualquer, manteve a decisão de seguir no mesmo curso. Já com o aluno de transferência interna, o ocorre o inverso e, para esses, a chance de evasão aumenta. Ele troca de curso, mantendo-se na IES. Nesse caso, pode-se inferir que esse aluno mostra ter mais dúvida quanto a sua trajetória acadêmica, apesar de optar por seguir os estudos na mesma instituição, mostrando ter algum vínculo com a

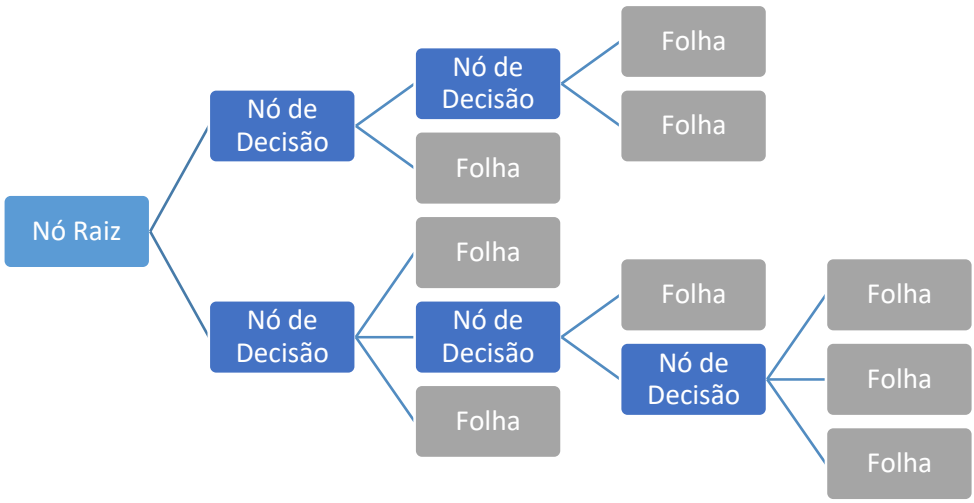
IES. A importância do tipo de ingresso aparece novamente no resultado da árvore de decisão e será novamente explorada neste estudo.

Alunos do PROUNI também aparecem com maior chance de evasão, indicando serem mais propensos a evadir, tudo mais constante. Mas curiosamente, alunos vindos de escolas públicas aparecem com chance mais baixa de evasão. Dois atributos parecem não ter nenhuma relação com a classe investigada, a distância em km da residência a instituição e o número de anos desde que saiu da escola. Com valores muito próximos a 1, pode-se dizer que esses atributos são insignificantes para explicar a classe, de acordo com a amostra utilizada. Por fim, o curso de fonoaudiologia (FO) aparece como um *outlier*, com grande chance de evasão: 30 alunos evadidos para cada formado, no entanto trata-se de um vício amostral devido ao baixo número de estudantes desse curso (25) na amostra.

4.1.3. Análise do modelo de árvore de decisão

O algoritmo de árvore de decisão classifica os dados a partir de regras (nós) que os dividem em ramos – essas divisões também são conhecidas como *splits*. Na ponta de cada ramo, novas regras podem ser aplicadas, criando novos nós que levam a novos *splits*. Quando um nó não é mais subdividido, é chamado de folha. O primeiro nó é conhecido como nó raiz e os nós intermediários, nós de decisão. A figura 4.1 ilustra uma árvore de decisão genérica para fins de exemplo.

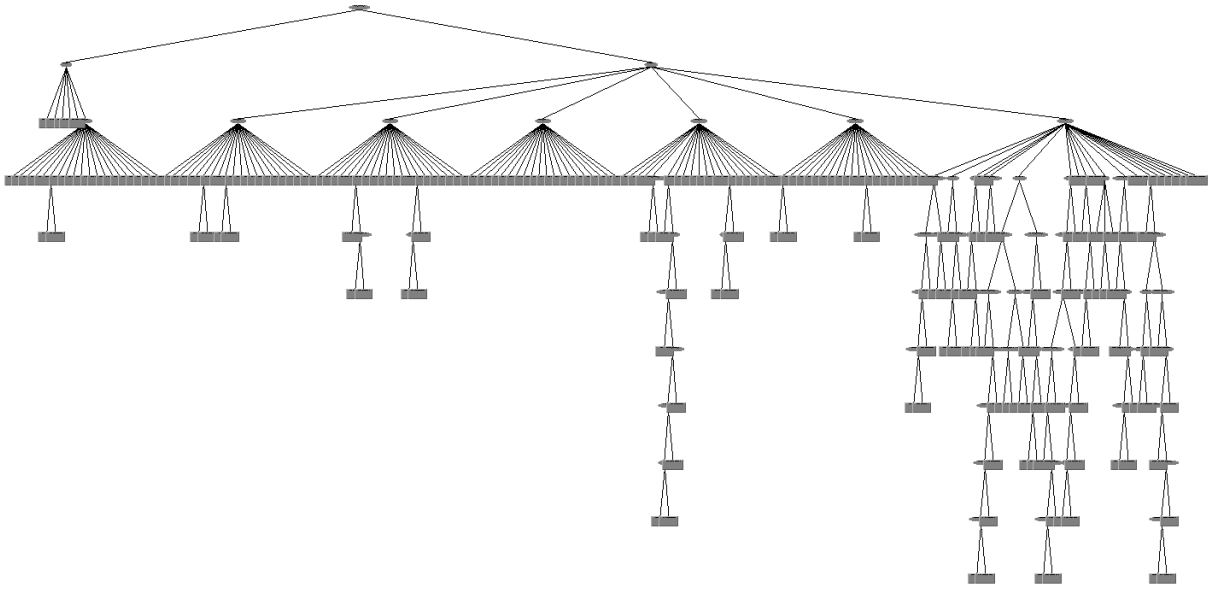
Figura 4.1. – Exemplo de estrutura de árvore de decisão.



Os ramos podem ter tamanhos diferentes e o nós podem ser divididos em dois ou mais grupos. A ideia é que, uma vez criada a estrutura da árvore, novos dados, não classificados perfaçam o caminho dos ramos de acordo com as regras impostas em cada nó. Quando chega na folha, o dado é identificado como pertencente a uma das classes.

O tamanho total da árvore calculada foi de 302 nós e folhas, indo de 3 a 11 níveis do nó raiz à folha dependendo do ramo seguido. A figura 4.2 ilustra a forma da árvore gerada pelo algoritmo.

Figura 4.2. – Estrutura da árvore de decisão calculada com algoritmo REPTree.



Um dos possíveis problemas que surgem com a aplicação de modelos de árvore de decisão é o chamado *overfiting*, quando a árvore gerada possui uma estrutura complexa, com muitos níveis, bastante específicos. Embora possa explicar bem a amostra na qual foi construído, pode não ter um bom poder preditivo, pois alguns dos nós criados não guardam relação com fenômenos reais, sendo apenas fruto de idiossincrasias da amostra em que foram baseados. No caso do nó de decisão de terceiro nível mais direita da figura 4.2, por exemplo, o desbalanceamento do seu tamanho pode indicar ser o caso.

Exibir a árvore e todos os seus nós pode ser complexo, mas examinar seus primeiros nós de decisão pode dar alguns *insights* para a gestão da instituição. Os primeiros *splits* indicam variáveis mais determinantes para a classificação da instância de acordo com o algoritmo, portanto, deve-se prestar mais atenção a elas. No primeiro *split*, o nó raiz é a média das notas do segundo período, dividindo em dois ramos, média maior ou menor do que 3,9 (numa escala de 0 a 10). Os nós seguintes, em ambos os ramos, o são o tipo de ingresso na instituição. Em seguida, no terceiro nível, temos os nós que dividem os alunos nos cursos. Portanto, vê-se aqui também que a média do segundo período, a forma de ingresso e o curso são atributos importantes na determinação da classe do aluno, corroborando os resultados encontrados no modelo construído via regressão logística.

O *split* dos alunos com média menor do que 3,9 está representado na figura 4.2 no primeiro nó de decisão abaixo do nó raiz à esquerda. Vê-se que a quantidade de nós a partir desse *split* é bem menor, pois esse fator apenas é um forte determinante da evasão. Essa nota é sensivelmente mais baixa que a nota mínima de corte para aprovação nas disciplinas, 5 pontos de 10 na média final, mas de maneira geral, 83% dos alunos que obtiveram uma média menor do que 3,9 no segundo período evadiram. A figura 4.3 indica os dois primeiros nós da árvore e as respectivas probabilidades de evasão:

Figura 4.3. – Estrutura parcial da árvore de decisão: primeiros dois níveis.

Variável: media_periodo_2

Split	P. Evasão
<3,9	83%
>=3,9	28%

<3,9 >=3,9

Variável: ingresso

Split	P. Evasão
ENEM	100%
PORTADOR DE DIPLOMA	79%
PROUNI	100%
TRANSFERENCIA EXTERNA	70%
TRANSFERENCIA INTERNA	80%
VESTIBULAR	92%

Variável: ingresso

Split	P. Evasão
ENEM	50%
PORTADOR DE DIPLOMA	14%
PROUNI	42%
TRANSFERENCIA EXTERNA	15%
TRANSFERENCIA INTERNA	38%
VESTIBULAR	35%

Analisando o lado esquerdo, o nó subsequente à média do segundo semestre é a forma de entrada, onde o modelo retorna evasão para todas as modalidades (são todos folhas) exceto para o vestibular (nó de decisão) – caso essa seja a modalidade de entrada, ele vai para um terceiro *split*, de acordo com o curso. O modelo então deduz que para todos os cursos o aluno evadiria, exceto para o curso de educação física. Neste, ele abre novamente mais um *split* onde indica que, caso o desconto médio do segundo período seja inferior a 15% o aluno concluirá o curso e evadirá caso contrário. É preciso um estudo mais aprofundado – com uma amostra maior – para concluir que, nessas condições, o nível do desconto realmente seja um fator determinante para conclusão do curso e não apenas uma questão de *overfitting*. Esse caso pode ser um *outlier*, pois para todos os demais casos o critério de nota do primeiro *split* foi suficiente para o modelo classificar o aluno.

Do lado direito da árvore, após o *split* inicial da nota do segundo período maior ou igual a 3,9, novamente é o critério de ingresso que se coloca como a variável mais importante para a determinação da classe, mas diferentemente do que lado esquerdo, deste lado as formas de ingresso não são folhas, mas sim nós de decisão, dependentes dos cursos. Como exibe a figura 4.3, as probabilidades de evasão aqui variam de 50% (ENEM) a 14% (portador de diploma). A tabela 4.7 compreende o próximo *split* e relaciona as probabilidades de evasão para cada tipo de ingresso e curso. Antes porém, a tabela 4.6 reproduz as informações dispostas no capitulo 3, que associa os cursos às suas respectivas siglas e denota o tamanho da amostra, para facilitar a leitura da tabela 4.7.

Tabela 4.6 – Reprodução da tabela 3.1: Tamanho da amostra por curso e duração.

Curso	Sigla	Duração (sem.)	Amostra
Psicologia	PS	10	444
Enfermagem	EN	10	686
Educacao Física	EF	8	864
Biologia	BI	8	267
Administração	AA	8	397
Fisioterapia	FS	10	91
Tecnólogo em Gestão de Recursos Humanos	TR	8	354
Ciências Contábeis	AC	4	183
Engenharia Ambiental e Sanitária	EA	8	124
Nutrição	NU	10	132
Farmácia	FM	8	337
Fonoaudiologia	FO	8	25
Tecnólogo em Estética e Cosmética	TE	10	97
Tecnólogo em Gestão de Marketing	TM	4	33
Tecnólogo em Processos Gerenciais	TG	4	6
Total			4040

Tabela 4.7. – Probabilidade de evasão por ingresso e curso (media_periodo_2 >= 3,9)

		Ingresso					
	Prob. Evasão Ingresso/Curso	ENEM	PORTADOR DE DIPLOMA	PROUNI	TRANSFERENCIA EXTERNA	TRANSFERENCIA INTERNA	VESTIBULAR
Curso	PS	68%	33%	100%	25%	75%	48%
	EN	76%	10%	100%	15%	17%	37%
	EF	40%	7%	16%	17%	38%	37%
	BI	43%	12%	100%	17%	29%	33%
	AA	60%	25%	14%	10%	50%	37%
	FS	100%	25%		21%	100%	46%
	TR	10%	11%	17%	7%	16%	8%
	AC	71%	18%		11%	50%	44%
	EA	86%	23%		38%		57%
	NU	64%	25%		37%	50%	47%
	FM	58%	6%	100%	5%	50%	48%
	TE	11%	0%		11%	67%	16%
	TM	0%	0%		0%	33%	9%
	TG					0%	75%
	FO	100%	100%		100%	50%	100%

As células da matriz na tabela 4.7 foram pintadas em escala de cores do verde (mais baixa) para o vermelho (mais alta), para facilitar a identificação das combinações que levam a maior risco de evasão. As células em branco representam casos onde não houve amostra da combinação. Casos extremos, onde há 100% ou 0% são fruto de amostras pequenas, com poucas observações. Alguns padrões podem ser vistos. Alunos que ingressaram via transferência externa ou já são portadores de diploma tem chance consideravelmente mais baixa de evasão, independente do curso, exceto fonoaudiologia, que parece ser um *outlier*. Os alunos ingressantes via vestibular interno tem probabilidade alta de evasão nos cursos de graduação, mas baixa para os cursos técnicos, exceto no curso de tecnólogo em processos gerenciais. Padrão similar aparece no ENEM, com probabilidades altas de evasão para cursos de graduação e baixas para cursos presenciais. Essa informação parece indicar um tipo de viés de seleção no processo de captação de alunos por esses canais, sugerindo que o processo atual funciona bem para alunos de cursos técnicos, mas não para graduação.

Após os cursos, a árvore de decisão cria outros nós mais específicos cujos detalhes não serão cobertos aqui – a insuficiência de amostras mais robustas pode ter ocasionado problemas de *overfitting* para os demais nós subsequentes.

4.2. Projeções (Scoring)

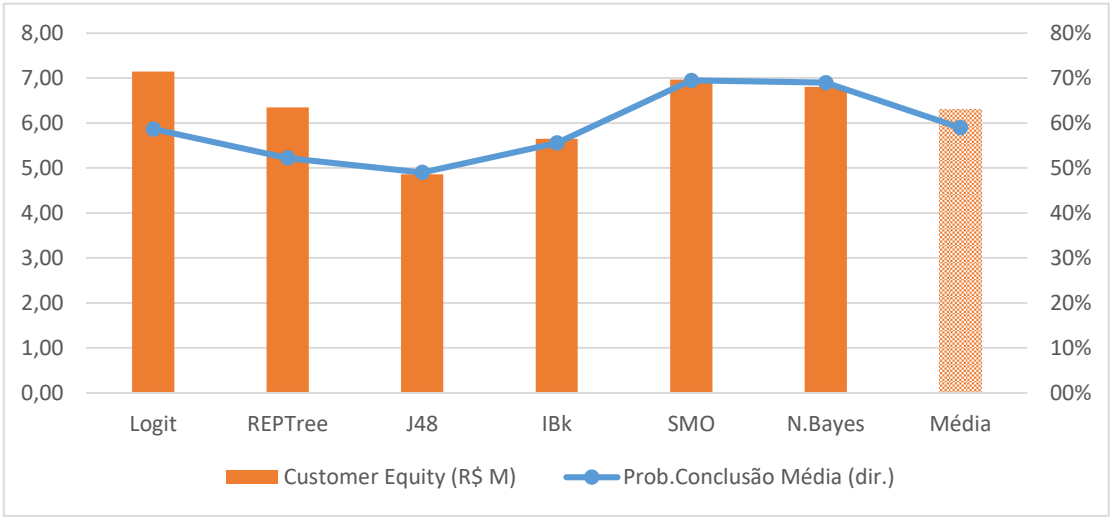
O objetivo de modelagens é entender um fenômeno para ser capaz de, a partir de certos elementos relacionados, prever o comportamento futuro da variável de interesse. Uma vez gerados os modelos, eles foram utilizados para fazer o *scoring* da base de alunos ativos. Individualmente, os resultados das previsões dos modelos, mostrados na tabela 4.8, variaram significativamente. A média dos resultados indica 41% de probabilidade de evasão média da base, o que significa dizer que dos cerca de 1.500 alunos ao final do segundo e terceiro períodos, aproximadamente 550 não devem concluir o curso, tudo mais constante. Essa projeção de evasão é próxima da obtida por Balaniuk et al (2011) cujo estudo apontou uma evasão futura estimada em 38,5% para uma amostra de alunos do segundo período.

O *customer equity* da amostra (soma dos *lifetime values* individuais) indica um valor de aproximadamente R$ 6,3 milhões. Todavia, esse resultado pode ser otimista, pois considerando os resultados analisados na seção 4.1.1, viu-se que os modelos foram sistematicamente conservadores em suas capacidades de prever a evasão: quando classificados como evadidos em geral os modelos acertaram, porém muitos casos de alunos com essa classe não foram identificados.

Tabela 4.8. – Probabilidade de evasão média e *customer equity* por modelo.

	Logit	REPTree	J48	IBk	SMO	N.Bayes	Média
Prob.Evasão Média	41.4%	47.8%	51.0%	44.4%	30.5%	31.0%	**41.0%**
Customer Equity (R$ M)	7.14	6.35	4.86	5.65	6.97	6.80	**6.29**

Figura 4.4. – Gráfico de probabilidade de conclusão e *customer equity* por modelo.



Naturalmente, quanto maior a taxa de conclusão média (probabilidade complementar a da evasão) indicada pelo modelo, maior seu *lifetime value*. A figura 4.4 ilustra relação direta dessas duas variáveis. No entanto, comparando os resultados dos diferentes modelos temos casos de taxas de conclusão diferentes

como valores próximos. Pela previsão da regressão logística, cerca 60% dos alunos da amostra concluirão seu curso, 10% a menos que nos modelos de *naive bayes* e *support vector machine*. No entanto o *lifetime value* do primeiro é o mais alto de todos. A diferença se deve ao valor dos alunos que foram classificados como futuros concluintes. A regressão logística classificou, na média, mais alunos valiosos como concluintes do que nos demais modelos.

5.
Conclusão

Este estudo buscou usar dados comuns ao ambiente de negócios estudado, usualmente disponíveis nos sistemas gerenciais utilizados por uma instituição de ensino superior para criar uma ferramenta de apoio ao marketing de relacionamento. Baseado na premissa de que o comportamento observado de estudantes no passado pode ser útil para entender o comportamento futuro de um estudante atual, o objetivo primário deste estudo foi estimar o risco de evasão de alunos ativos de uma IES privada, utilizando dados acadêmicos, financeiros e demográficos.

Para isso, foram percorridas as seguintes etapas, definidas como objetivos intermediários: (i) coleta dos dados dos sistemas da IES, transformação das variáveis nos atributos relevantes para o problema e consolidação num banco de dados próprio para leitura no software de *machine learning* WEKA; (ii) construção e teste de modelos preditivos de classificação no software WEKA a partir dos algoritmos de regressão logística, árvore de decisão (*REPTree*), J48, *k-nearest neighbors*, *support vector machines* e *naive bayes*; (iii) análise comparativa dos resultados dos modelos calculados; (iv) estimação da probabilidade de evasão de uma amostra representativa de alunos ativos a partir dos modelos calculados.

Do ponto de vista gerencial, os resultados das previsões permitem à instituição ter uma lista classificatória dos alunos em termos de risco de evasão e de valor econômico (*lifetime value*). Assim, a IES tem à disposição um *ranking* de risco, a partir do qual pode montar uma estratégia para decidir onde concentrar seus esforços de retenção. A robustez dos resultados depende diretamente do tamanho e relevância da amostra. Considerando deficiências dos sistemas de informação para recuperação de dados históricos e o próprio tamanho da instituição, o estudo buscou utilizar o maior número possível de alunos para desenvolver os modelos e selecionou, dentre todas informações existentes para as quais era possível obter valores históricos, os atributos mais relevantes. Para utilização prática, é importante que os modelos sejam continuamente alimentados com novas informações a cada novo semestre. Uma característica fascinante de algoritmos de aprendizado de máquinas é que, com tempo, espera-se fiquem cada vez mais robustos. O princípio

metodológico adotado neste trabalho também pode ser sofisticado com mais atributos na medida em que esses estejam disponíveis.

Ainda assim, alguns *insights* apareceram na análise dos dados. De todos os atributos, a média de notas – variável chave do desempenho acadêmico – em particular do segundo período, aparece como fator fundamental para determinar a futura evasão do aluno. Este resultado era esperado e está em linha com a principal teoria do abandono estudantil, de Tinto (1993), que vê na ausência de comprometimento acadêmico um dos dois principais vetores para a evasão, juntamente com a ausência de integração social.

A forma de ingresso na instituição também possui papel importante na determinação da evasão. Alunos ingressantes pelo ENEM possuem chance consideravelmente mais alta de evasão que os demais modos. É possível que este achado também tenha ressonância na teoria da integração de Tinto (1993). Pelo ENEM o aluno faz uma única prova, não vinculada a nenhuma instituição em particular e, posteriormente tenta ingressar nas IES que o aceitarem dada a sua nota. Já quando o aluno faz o vestibular interno da instituição, por exemplo, ele já demonstra, antes mesmo de conhecer seu potencial de acadêmico de ingresso (sua nota), interesse na instituição. Esta pode ser uma *proxy* para intenção de integração social na instituição. Outro achado que corrobora a visão de Tinto (1993) é que alunos vindos de transferência externa tem chance bem mais baixa de evasão. Neste caso, o aluno está mudando de instituição, mas mantendo a decisão de continuar naquele curso, o que demonstra certo grau de comprometimento acadêmico. Da mesma forma ocorre com alunos que já possuem diploma: estes já passaram por todo o processo acadêmico e voltaram a estudar, também demonstrando terem, tudo mais constante, um grau de comprometimento acadêmico maior do que o observado na amostra de alunos do ENEM.

Em outro exemplo de aderência dos resultados à teoria de Tinto (1993), a transferência interna aparece com chance alta de evasão: neste caso o aluno muda o curso, mas opta por se manter na instituição, mostrando que há integração social do aluno com a IES, ainda que não tenha havido a devida integração acadêmica – razão para o abandono do primeiro curso. Segundo Tinto (1993), uma excessiva integração social com baixa integração acadêmica também favorece a evasão, caso

em que um aluno prioriza atividades sociais em detrimento dos estudos. Enquadram-se neste caso alunos que, embora não se identifiquem com o curso, tenham amigos e uma interação social intensa na instituição e podem ter dificuldades em abandoná-la num primeiro momento. Assim, tentam outro curso na mesma instituição, resultando em uma evasão posteriormente.

Outro resultado com potencial impacto gerencial é o papel dos descontos de mensalidade. A regressão logística encontrou maior propensão à evasão quanto maior o desconto dado no semestre inicial. Pode estar havendo aí um caso de viés de seleção em que, a título de aumento da captação, descontos maiores são concedidos numa disputa para atrair mais alunos, porém o aumento da base de alunos ocorre em detrimento do nível de comprometimento do estudante com a IES. Além disso, a redução do nível de desconto do primeiro para o segundo período também aumenta as chances de evasão. Portanto uma vez que um certo patamar de desconto foi estabelecido, sua redução implica efeito negativo na probabilidade de retenção, embora o nível do desconto em si não seja tão significativo.

5.1. Oportunidades para futuras aplicações

O advento dos cursos de ensino a distância (EaD) cria um ambiente de negócios bastante propício à aplicação de ferramentas de *machine learning* para marketing de relacionamento. Ambientes virtuais são ricos em geração de dados e cada "passo" dado dentro desses sistemas online podem ser medidos e analisados. Além disso, com a relativa frieza de um contato puramente virtual com a instituição de ensino pode ser mitigada com um tratamento individualizado, reconhecendo as características e histórico de relacionamento de cada aluno. Além disso, cursos de EaD tem como apelo para as instituições a grande escalabilidade de custos, podendo atender a um grande número de alunos com menos recursos do que num curso presencial.

A metodologia aqui utilizada representa uma ferramenta absolutamente escalável que, na prática funciona melhor quanto mais instâncias existem, e que tem como output um identificador único para cada aluno baseado nos seus próprios dados. Portanto, este segmento de ensino representa uma oportunidade excelente

para aplicação de metodologias de *machine learning*, não apenas para classificação de evasão, mas também podendo ser exploradas em outras questões do negócio.

Referências bibliográficas

ADACHI, A. A. C. T. **Evasão e evadidos nos cursos de graduação da Universidade Federal de Minas Gerais**. 30 jan. 2009. 214 f. Dissertação – Universidade Federal de Minas Gerais. Belo Horizonte/MG, 30 jan. 2009.

AHER, S. B.; LOBO, L. M. R. J**. Data mining in educational system using Weka**. IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT), p. 20-25, 2011.

AL IMRAN, M. A. **A study on Amazon: information systems, business strategies and e-CRM**, 2014.

ALTMAN, N. S. **An introduction to kernel and nearest-neighbor nonparametric regression**. The American Statistician, 46.3, 175-185, 1992.

AULCK, L. et al. **Predicting Student Dropout in Higher Education**. ar Xiv preprint arXiv:1606.06364, 2016.

BALANIUK, R. et al. **Predicting evasion candidates in higher education institutions.** In International Conference on Model and Data Engineering, p. 143-151, Springer Berlin Heidelberg, setembro, 2011.

BARNES, J. B. **Secrets of Customer Relationship Management.** New York: McGraw-Hill, 2001.

BEAN, J. P.; METZNER, B. S. **A Conceptual Model of Nontraditional Undergraduate Student Attrition**. Review of Educational Research, 55, 485-540, 1985.

BERGAMO, F.V. M.; GIULIANI, A. C. **A lealdade do estudante baseada na qualidade do relacionamento: uma análise em instituições de ensino superior**. XXXIII Encontro Nacional da Associação Nacional dos Programas de Pós-Graduação em Administração–ANPAD, 2009.

BERGER, Paul D.; NASR, N. I. **Customer lifetime value: Marketing models and applications**. Journal of interactive marketing, v. 12, n. 1, p. 17-30, 1998.

BERRY, M.; LINOFF, G. **Mastering Data Mining: The Art and Science of Customer Relationship Management**. John Wiley & Sons, 2000.

BLATTBERG, R. C.; DEIGHTON, J. **Managing Marketing by the Customer Equity Test**. Harvard Business Review, 74.4, p. 136-144, 1996.

BORGES, S. M. **Fatores determinantes da evasão escolar no ensino superior**: o estudo de caso DOILES/ ULBRA de Itumbiara. 77f. Dissertação (Mestrado Profissional em Desenvolvimento Regional) – Faculdades Alves Faria, 2011.

CORTES, C., VAPNIK, V. **Support-vector networks**. Machine learning, 20.3, p. 273-297, 1995.

CRAMER, J. S. **The origins of logistic regression**. Tinbergen Institute Discussion Paper, 119/4, 2002.

DELEN, D. **A comparative analysis of machine learning techniques for student retention management**. Decision Support Systems, v. 49, n. 4, p. 498-506, 2010.

DICK, A. S., BASU, K. **Customer loyalty: toward an integrated conceptual framework**. Journal of the academy of marketing science, 22.2, P. 99-113, 1994.

DURKHEIM, E. **Suicide**. J. Spaulding & G. Simpson, trans. Glencoe: The Free Press, 1961.

FAWCETT, T. **An introduction to ROC analysis**. Pattern recognition letters, 27.8, p. 861-874, 2006.

FILIATRAULT, P.; LAPIERRE, J. **Managing Business-to-Business Marketing Relationships in Consulting Engineering Firms**. Industrial Marketing Management, 26.2, pp. 213-222, 1997

FRADKIN, D.; MUCHNIK, I. **Support vector machines for classification**. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 70, p. 13-20, 2006.

GERBA, R. T. **Análise da evasão de alunos nos cursos de Licenciatura: Estudo de caso no Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina.** 17 set. 2014. 157 f. Dissertação – Universidade Federal de Santa Catarina. Florianópolis/SC, 17 set. 2014.

GRUBB, W. Norton. **Dropouts, Spells of Time and Credits in Postsecondary Education: Evidence from Longitudinal Surveys**. Economics of Education Review, 8 (1), 49-67, 1989.

HALL, M., et al. **The WEKA data mining software: an update**. ACM SIGKDD explorations newsletter, 11.1, p. 10-18, 2009.

HAN, J.; KAMBER, M. **Data mining: Concepts and Techniques**. 2. ed. [S.l.]: Morgan Kaufmann San Francisco, Calif, USA, 2006.

HAND, D J.; YU, K**. Idiot's Bayes – not so stupid after all?** International statistical review, 69.3, p. 385-398, 2001.

HENNIG-THURAU, T.; LANGER, M. F.; HANSEN, U. **Modeling and managing student loyalty: an approach based on the concept of relationship quality**. Journal of Service, 2001.

HOGAN, J.E.; KATHERINE, N.L.; BARAK, L. **What is the true value of a lost customer?** Journal of Service Research, Vol. 5 No. 3, pp. 196-208, 2003.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior**. Brasília: INEP, 2016. Disponível em: http://portal.inep.gov.br/web/censo-da-educacao-superior. Acesso em: 18 nov. 2015.

JAIN, D.; SINGH, S. S. **Customer lifetime value research in marketing**: A review and future directions. Journal of interactive marketing, 16.2, p. 34-46, 2002.

KOTLER, P. **Administração de marketing: análise, planejamento, implementação e controle**. Atlas, 1994.

KOTSIANTIS, S. B. **Supervised machine learning: A review of classification techniques**. Informatica, 31.3, p. 249-269, 2007.

LEE, W.; STOLFO, S. J.; MOK, K. W. **A data mining framework for building intrusion detection models**. Security and Privacy. Proceedings of the 1999 IEEE Symposium on. IEEE, p. 120-132, 1999.

LEE-KELLEY, L.; GILBERT, D.; MANNICOM, R. **How e-CRM can enhance customer loyalty.** Marketing Intelligence & Planning, Vol. 21 No. 4, pp. 239-48, 2003.

LOBO, M. B. C. M. **Panorama da evasão no ensino superior brasileiro:** aspectos gerais das causas e soluções. ABMES Cadernos. Brasília, set./dez, 2012.

LUCHESA, C. J.; MACHADO, C. R. **A Margem de Contribuição na Gestão de Instituições de Ensino Superior–IES**. Teoria e Prática da Educação, v. 14, n. 2, p. 113-122, 2012.

MAHISHI, A. **Customer Lifetime Value** – Not Just a Marketing Metric. Tata Consultancy Services, 2014. Disponível em: http://www.tcs.com/SiteCollectionDocuments/White-Papers/Customer-Lifetime-Value-Not-Just-a-Marketing-Metric-1214-1.pdf. Acesso em: 1 mar. 2017.

PENG, C. Y. J.; LEE, K. L.; INGERSOLL, G. M. **An introduction to logistic regression analysis and reporting**. The journal of educational research, 96.1, p. 3-14, 2002.

PEPPERS AND ROGERS GROUP. **Roper Starch Worldwide survey**. Setembro, 2000.

PEPPERS, D.; ROGERS, M. **Managing Customer Relationships***:* A Strategic Framework. Nova Jersey: John Wiley & Sons, Inc, 2011.

PEREIRA FILHO, E. **Compromisso com o graduar-se, com a instituição e com o curso**: estrutura fatorial e relação com a evasão. 89 f. Dissertação (Mestrado em Educação) – Universidade Estadual de Campinas, São Paulo, 2012.

PRIM, A. L.; FÁVERO, J. D. **Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau**. Revista E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, n. Especial Educação, p. 53-72, 2013/2.

QUINLAN, J. R**. Improved use of continuous attributes in C4.5**. Journal of artificial intelligence research, 4, p. 77-90, 1996.

RAPP, S.; COLLINS, T. L. **The New Maximarketing**. McGraw-Hill, 1995.

RISH, I. **An empirical study of the naive Bayes classifier**. IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, n. 22, p. 41-46, IBM New York, 2001.

ROUSSEAU, Christiane. **How Google works**. Klein vignette *(www. kleinproject. org*), 2010.

SCHWARTZMAN, J. **O Financiamento das Instituições de Ensino Superior no Brasil**. Paper. Instituto de Estudos Avançados, Universidade de São Paulo, 2003.

SEIDMAN, Alan. **Retention Revisited**: R=E,Id+E&In,Iv. College and University, 71.4, p. 18-20, 1996.

SHARMA, T. C.; JAIN, M. **WEKA approach for comparative study of classification algorithm**. International Journal of Advanced Research in Computer and Communication Engineering, 2.4, p. 1925-1931, 2013.

SILVA FILHO, R. L. L. et al. **A evasão do ensino superior brasileiro.** Cadernos de Pesquisa, São Paulo, v. 37, n. 132, p. 641-659, 2007.

SILVA, H. R. B., ADEODATO, P. J. L. **A data mining approach for preventing undergraduate students retention**. In: The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012. p. 1-8.

THAKUR, R; SUMMEY, J. H. **Optimizing CRM: A framework for enhancing profitability and increasing lifetime value of customers**. Marketing Management Journal, 20.2, p. 140-151, 2010.

TIERNEY, William G. **An Anthropological Analysis of Student Participation in College**. Journal of Higher Education, 63.6, p. 603-618, 1992.

TINTO, V. **Dropout from higher education: a theoretical synthesis of recent research**. Review of Educational Research, 45, 89-125, 1975.

TINTO, V. **Leaving college**: Rethinking the causes and cures of student attrition. The University of Chicago Press, Chicago, 1987.

TINTO, V. **Leaving college**: Rethinking the causes and cures of student attrition. 2 ed. Chicago: University of Chicago Press, 1993.

VOGEL, V.; EVANSCHITZKY, H.; RAMASESHAN, B. **Customer equity drivers and future sales**. Journal of marketing, 72.6, p. 98-108, 2008.

WEST, D. **Neural network credit scoring models**. Computers & Operations Research, v. 27, n. 11, p. 1131-1152, 2000.

WILSON, J. R.; LORENZ, K. A. **Short History of the Logistic Regression Model**. In: Modeling Binary Correlated Responses using SAS, SPSS and R, p. 17-23, Springer International Publishing, 2015.

WITTEN, I. H.; FRANK, E.; HALL, M. A**. Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. Burlington: Morgan Kaufmann, 2011.

ZEKIC-SUSAC, M.; SARLIJA, N.; BENSIC, M. **Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models.** Em: Information Technology Interfaces, 2004. 26th International Conference on. IEEE, p. 265-270, 2004.