JOÃO GABRIEL CORRÊA KRÜGER

UMA ABORDAGEM DE APRENDIZAGEM DE MÁQUINA EXPLICÁVEL PARA PREVISÃO DE EVASÃO ESTUDANTIL

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná (PUCPR) como requisito parcial para obtenção do título de Mestre em Informática.

Curitiba
2022

JOÃO GABRIEL CORRÊA KRÜGER

UMA ABORDAGEM DE APRENDIZAGEM DE MÁQUINA EXPLICÁVEL PARA PREVISÃO DE EVASÃO ESTUDANTIL

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná (PUCPR) como requisito parcial para obtenção do título de Mestre em Informática.

Área de concentração: Ciência de Dados

Orientador: Jean Paul Barddal
Co-orientador: Alceu de Souza Britto Junior

Curitiba
2022

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Pamela Travassos de Freitas CRB/9 1960

Krüger, João Gabriel Corrêa
K94u 2022 Uma abordagem de aprendizagem de máquina explicável para previsão de evasão estudantil / João Gabriel Corrêa Krüger ; orientador: Jean Paul Barddal ; Co-orientador: Alceu de Souza Britto Junior. – 2022.
100 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2022
Bibliografia: f. 80-87

1. Informática. 2. Aprendizado do computador. 3. Algoritmos. 4. Evasão escolar. 5. Previsão (Lógica). I. Barddal, Jean Paul. II. Britto Junior, Alceu de Souza. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

53-2022

DECLARAÇÃO

Declaro para os devidos fins que o aluno **JOÃO GABRIEL CORRÊA KRÜGER**, defendeu sua dissertação de Mestrado intitulada "**Uma Abordagem de Aprendizagem de Máquina Explicável para Previsão de Evasão Estudantil",** na área de concentração Ciência da Computação, no dia 10 de junho de 2022, no qual foi aprovado.

Declaro ainda que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade, firmo a presente declaração.

Curitiba, 05 de agosto de 2022.

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná

Sometimes the best way to solve your own problems is to help someone else.
Uncle Iroh

Agradecimentos

Aos meus pais, Fábio e Liege, e minha irmã Mariana, que em nenhum momento deixaram de me apoiar. Sem vocês, não estaria aqui. Me esforço todo dia a acordar e me tornar uma pessoa melhor, na esperança de um dia me tornar como vocês. São meus exemplos éticos e profissionais. A paciência, a presença, o amor, o carinho e a sabedoria de vocês me guiaram até aqui, os quais carregarei comigo a vida toda.

Aos meus amigos, em especial Paulo, Michel e Marina. Seus conselhos, paciência e carinho são de valor indescritível. As conversas, os incentivos e os bons momentos, mesmo que as vezes a distância, são seus presentes mais valiosos. Espero continuar caminhando próximo a vocês pelo resto da minha vida.

Aos meus familiares, que mesmo longe sempre me apoiaram, cuidaram, amaram e respeitaram. A distância é grande mas o carinho também.

Aos meu orientadores, Prof. Jean e Prof. Alceu. Suas contribuições para minha formação vão além deste trabalho. Ambos são profissionais que tomo como exemplo para mim. A disponibilidade, o esforço e a dedicação de vocês é inigualável. Espero ter a oportunidade de trabalharmos juntos no futuro novamente.

Aos professores Fabrício, Júlio e Albino. Suas contribuições para esse trabalho foram de grande valia. Seus conselhos e sugestões são altamente respeitados por mim, e tiveram um impacto positivo neste projeto.

Aos meus ex-colegas no Grupo Marista. Foram 2 anos de muito aprendizado e trabalho. O impacto causado na vida de tantos estudantes é algo que todos carregam com orgulho. Desejo muito sucesso na carreira profissional de vocês.

Aos demais bolsistas pela companhia, apoio, conversas e sugestões. Espero que vocês tenham muito sucesso na carreira adiante.

Aos Professores e aos colegas do PPGIa, que mesmo a distância contribuíram para minha formação.

Ao Grupo Marista pelo apoio estrutural e financeiro para o desenvolvimento deste projeto.

Muito obrigado.

Resumo

A evasão escolar é um problema dentro da área da Educação com grande impacto social e econômico, e portanto, é um tópico de grande interesse a ser estudado. Consequentemente, a aplicação de técnicas como a previsão de evasão usando aprendizagem de máquina possui potencial para mitigar esse problema. Entretanto, tão importante quanto a previsão de uma evasão estudantil é o entendimento dos motivos que a ela provocaram, tópico que não é amplamente abordado pela literatura devido às técnicas serem mais recentes. Este trabalho apresenta o levantamento de um conjuntos de dados a partir das informações de estudantes do Grupo Marista, o treino de modelos de classificação para distinguir alunos evasores e não evasores e a subsequente aplicação de técnicas de interpretação de modelos no contexto de evasão estudantil. Considerando métricas adequadas a conjuntos de dados desbalanceados, foi constatada a eficácia dos modelos desenvolvidos na previsão de evasão em diversos segmentos de ensino e momentos do ano escolar. Os modelos desenvolvidos apresentaram melhores desempenhos em estudantes do Ensino Médio e Infantil, com valores respectivos de AUC PR de 79.71% e 82.86%, quando comparados aos do Ensino Fundamental anos iniciais e finais, os quais obtiveram 66.84% e 68.25% respectivamente. Os modelos também apresentaram melhores resultados quanto mais tarde no ano letivo a previsão foi feita, sendo os melhores resultados no terceiro trimestre, o qual apresentou AUC PR de 92.33%. Este trabalho também ilustra, por meio de técnicas de explicação de modelo, que notas e indicadores socio-econômicos são fatores chave que levam ou refletem a evasão dos estudantes e apresenta maneiras de representar esse problema para times não técnicos.

Palavras-chave: Previsão de Evasão; Educação; Aprendizagem de Máquina; Interpretação de Modelos.

Abstract

School dropout is a problem of great social and economic importance inside of the Education field, and therefore, is a topic of great interest to be studied. Consequently, the application techniques such as machine learning-powered student dropout prediction has potential to mitigate this problem. However, as important as predicting a student dropout is understanding the reasons that led to it, topic which isn't largely approached in the literature due to the techniques being recent. This dissertation presents the creation process of a dataset from Grupo Marista's students data, the training of classification models to distinguish between dropout and non dropout students and the subsequent application of model interpretation techniques in the student dropout context. Considering adequate metrics to evaluate imbalanced datasets, it was verified the efficacy of the developed models to predict student dropout in many different education segments and moments of the academic year. The developed models presented better results in high school and preschool students, presenting respective AUC PR scores of 79.71% and 82.86%, while compared to elementary and middle school students, which presented 66.84% and 68.25% respectively. The models also presented better results the later in the school year, with the third quarter presenting the best results, in which presented AUC PR of 92.33%. This dissertation also illustrates, by using model explaining techniques, that grades and socio-economic factors are key factors that lead or reflect students quitting their studies and presents manners of representing this problem to non-techincal teams.

Keywords: Dropout prediction; Education; Machine Learning; Model Interpretation.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com o aumento do uso de tecnologias em múltiplas áreas e empregos, criam-se muitas expectativas sobre o futuro da economia e o mercado de trabalho. Embora haja muita incerteza se as mudanças gerarão impactos positivos ou negativos na sociedade como um todo, é consenso que a tecnologia impacta na maneira em que tarefas são executadas. Empregos, mercados e possibilidades novas são frutos do advento da adesão à tecnologia em processos (RA et al., 2019).

Em mercados onde existe a fidelização de clientes, como saúde, educação ou estética, uma grande parte da fonte de renda prevista em seu modelo de negócio vem do retorno e pagamento frequentes de clientes. As empresas inseridas nesses mercados, os quais são altamente competitivos, necessitam de constante inovação e investimento para se manterem relevantes (MITCHELL; COLES, 2004).

Dentro dos setores do mercado que sofrem impactos diretos decorrentes da tecnologia encontra-se o setor da educação. O setor da educação, em especial o da educação privada, é altamente competitivo e pode ser impactado de diversas maneiras pela tecnologia. A tarefa de integrar tecnologias ao ensino, seja para auxiliar na aprendizagem ou na maneira que é feita a interação com os alunos, não é nova: existem esforços desde o século XVIII (SCHINDLER et al., 2017) para integrar novas tecnologias ao ensino. Conforme os anos passaram, tecnologias como o vídeo e a internet têm estado cada vez mais presentes na vida do aluno moderno (WESTERA, 2015).

Essa mudança, embora apresente divergência com relação a abordagens mais clássicas, como o uso exclusivo de livros e cadernos, pode possuir impactos positivos na experiência dos alunos, possibilitando melhorias não antes possíveis em sala de aula por meio do uso dos dados gerados(HEFLIN; SHEWMAKER; NGUYEN, 2017).

O uso de plataformas digitais para controle de alunos e ensino a distância permitiu a possibilidade de fazer um maior acompanhamento dos alunos em sua jornada estudantil,

seja por coletar dados de tempo gasto em estudos, questões, atenção nas aulas ou ainda fazer um melhor controle de notas e presença dos alunos (KERR, 2015).

Entretanto, embora existam benefícios advindos da tecnologia, nem todos os alunos e professores tiveram uma boa adaptação à adoção de tecnologias novas no ensino (HEFLIN; SHEWMAKER; NGUYEN, 2017). Além disso, alunos com problemas pessoais podem ter dificuldades no aprendizado, levando à desistência (BARDACH et al., 2019). A situação financeira também pode ser de grande impacto na permanência na escola, principalmente em escolas pagas.

Levando em conta os diferentes fatores que podem levar à desistência, ou evasão, de um aluno pode-se afirmar que a complexidade nos casos de evasão estudantil é grande. A evasão estudantil possui um grande impacto na qualidade do ensino do país. Indicadores como taxa de alfabetização e nível de compreensão matemática são capazes de influenciar a vida futua do estudante (JENKINS et al., 2018). Levando em consideração a perspectiva de um bom futuro como fator motivador ao estudante, é notável a importância do estudo da evasão estudantil dentro do âmbito da sociedade.

Além do enorme impacto social, o problema de evasão possui um grande impacto para o mercado de educação privada, visto que é sua principal fonte de renda. Sendo assim, a possibilidade de identificar uma potencial evasão, ou suas causas, é de grande interesse econômico (MÁRQUEZ et al., 2016). Prever desistências pode ajudar com o controle de fluxo de caixa e atuar como ferramenta para indicar a saúde do negócio.

A previsão de *churn*, ou perda de clientes, é possível fazendo o uso de técnicas de aprendizado de máquina. Para tal, geralmente é necessária uma grande quantidade de dados históricos dos clientes. Levando em consideração a similaridade da evasão com *churn*, é possível fazer uso de técnicas semelhantes, utilizando conjuntos de dados estudantis, para prever casos de evasão (QUARTI; FIGINI; GIUDICI, 2009).

Além da previsão de estudantes evasores, por meio de técnicas de avaliação do modelo de aprendizado de máquina, é possível diagnosticar o que levaria o estudante à largar a instituição usando técnicas de explicação de modelos (LUNDBERG; LEE, 2017). O uso prévio de técnicas para explicação de modelos de aprendizagem de máquinas aplicados à evasão estudantil sejam desconhecidos durante o momento da escrita deste documento, ao contrário do contexto de *churn* (VILLARREAL et al., 2020; DUMITRACHE; NASTU; STANCU, 2020).

Embora a evasão escolar seja um problema similar a *churn*, os fatores motivantes e os comportamentos apresentados são diferentes e, portanto, podem elucidar potenciais problemas e melhorias nas técnicas de explicação de resultados levando assim a ganhos nas áreas de aprendizagem de máquina e educação.

Além disso, diagnósticos de motivos que levam (ou ao menos evidenciem) a evasão podem possibilitar melhorias no negócio e na qualidade de serviço oferecido por meio da atuação de profissionais da educação, os quais não possuem tempo hábil suficiente para monitorar a grande massa de alunos individualmente.

Portanto, levando em consideração a complexidade, a relevância das explicações de resultados de modelos, a importância econômica e social justifica-se a pesquisa de diferentes técnicas de aprendizado de máquina para previsão e sua explicação aplicados no contexto da evasão estudantil.

1.1 Objetivo

Tendo em mente o funcionamento geral do setor de educação privada, o objetivo deste trabalho é desenvolver modelos usando técnicas de aprendizagem de máquina para a tarefa de predição de evasão de alunos. Para tal, pretende-se construir e avaliar modelos indutores que, além da detecção automática de potencial evasão, forneçam uma explicação para o motivo ou forneçam indícios desta por meio de técnicas de interpretação de modelos.

Os objetivos específicos deste trabalho são:

- Criar uma base de dados usando os dados dos sistemas educacionais da Educação Básica privada do Grupo Marista;
- Gerar novas características a partir dos dados dos alunos e bases externas, como as do IBGE, para obter uma melhor classificação;
- Treinar classificadores para distinguir alunos entre evasores e não evasores com base nos dados coletados;
- Colaborar com novas características e abordagens para o problema de classificação de evasão; e
- Aplicar técnicas para explicação de modelos de aprendizado de máquina sob a ótica da evasão estudantil de maneira a ilustrar quais fatores possuem maior impacto.

Este trabalho não tem dentro de seu escopo o acompanhamento do uso dos dados dos modelos treinados por parte do Grupo Marista, tendo este caráter corporativo.

1.2 Hipóteses do Trabalho

Existem duas hipóteses a serem validadas neste trabalho:

- Indutores que facilitam a explanação dos resultados (caixa-branca) são competitivos para a predição de evasão de alunos quando comparados a indutores caixa preta;
- Informações complementares ao desempenho do aluno como dados financeiros, socioeconômicos e a localização de sua moradia contribuem positivamente para o desempenho de um modelo preditivo de evasão;

1.3 Contribuições Científicas e Tecnológicas

Este trabalho é uma pesquisa aplicada na área da educação, e portanto suas contribuições envolvem também esta área. Este trabalho contribui com diferentes abordagens para as etapas de coleta de dados na tarefa de classificação de alunos em evasores e não evasores. Contribui também com o levantamento de *features* novas para o uso em casos de classificação de evasão estudantil e a sua contribuição para a classificação final. A compreensão da participação e influência dos atributos partirá tanto do uso de representação de modelos explicáveis, em algoritmos de caixa branca, quanto do estudo uso de técnicas de explicação de modelos, no caso de modelos caixa-preta, técnicas que não são amplamente abordadas na literatura no contexto de evasão estudantil.

1.4 Publicações

Este trabalho conta com uma série de publicações derivadas, sendo elas:

- João Gabriel Corrêa Krüger, Alceu de Souza Britto Junior e Jean Paul Barddal. Uma Abordagem de Aprendizagem de Máquina Explicável para Previsão de Evasão Estudantil. Em: III Fórum de Programas de Pós-Graduação em Computação do Paraná. 2021.
- **(Aceito)** João Gabriel Corrêa Krüger, Jean Paul Barddal e Alceu De Souza Britto Junior. A Machine Learning Approach for School Dropout Prediction in Brazil. *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN*. 2022.
- **(Submetido)** João Gabriel Corrêa Krüger, Jean Paul Barddal, Alceu De Souza Britto Junior, Jarbas Tadeu Madril e Caroline Ferreira Costa Serqueira. An Interpretable Machine Learning Approach for Student Dropout Prediction. *International Journal of Artificial Intelligence in Education*. 2022.

1.5 Organização

Este trabalho está dividido da maneira a seguir: o Capítulo 2 trata de uma revisão bibliográfica de conceitos de *churn*, evasão estudantil, aprendizado de máquina e interpretação de modelos. O Capítulo 3 aborda o estado da arte em previsão de evasão estudantil e interpretação de modelos. O Capítulo 4 expõe e discute as abordagens propostas, a qual é a avaliada no Capítulo 5. Por fim, o Capítulo 6 apresenta as conclusões gerais deste trabalho.

Capítulo 2

Fundamentação teórica

A educação, a qual é um direito de todos os cidadãos do Brasil (BRASIL, 1996), é uma peça fundamental da máquina estatal. O Brasil possui, atualmente, um sistema de educação pública oferecida pelo estado e diversas instituições privadas que oferecem este serviço.

O estado brasileiro define as normas e regras para a educação, tanto privada quanto pública, em toda a federação. O sistema educacional brasileiro foi originalmente definido pela Lei de Diretrizes e Bases da Educação, ou lei n.º 9.394 de 1996 (BRASIL, 1996), e redigida constantemente desde então. A lei aborda todos os tópicos de como deve ser estruturado o sistema no país. A lei também indica quais assuntos devem ser abordados nas salas de aula por meio da Base Nacional Comum Curricular. Essa divisão é feita para evidenciar as habilidades exigidas dos alunos e o contexto no qual estão inseridos.

As responsabilidade sobre as definições de conteúdo são de cada unidade federativa, as quais são obrigadas a seguirem as linhas gerais da Base Nacional Comum Curricular, abordando livremente os assuntos ditados pelo Governo Federal.

2.1 Organização do Sistema de Educação Brasileira

Conforme definido na lei 9.394 de 1996, o sistema brasileiro de educação é organizado em diferentes segmentos educacionais, cada qual almejando uma parcela da população e um nível de informação diferente.

A educação básica, definida na Lei de Diretrizes e Bases da Educação, tem caráter obrigatório e pode ser divida em Educação Infantil, Ensino Fundamental e Ensino Médio.

A educação infantil é composta de alunos de até 5 anos de idade e tem como finalidade o desenvolvimento integral da criança de até 5 (cinco) anos, em seus aspectos físico, psicológico, intelectual e social, complementando a ação da família e da comunidade

(BRASIL, 1996).

O ensino fundamental tem duração de 9 anos e pode ser iniciado aos 6 anos de idade. Ele tem como objetivo (BRASIL, 1996):

- Desenvolver as capacidades de aprendizagem e cognitivas;
- Compreender o domínio da escrita, leitura e cálculo; e
- Abordar valores sociais, políticos, tecnológicos e naturais da sociedade.

O ensino fundamental é comumente dividido em anos iniciais e anos finais. Os anos iniciais compreendem os 5 primeiros anos de ensino, enquanto os anos finais compreendem os 4 anos finais.

O ensino médio tem duração de 3 anos e tem como finalidade (BRASIL, 1996):

- Consolidar conhecimentos vistos previamente no ensino fundamental;
- Possibilitar o prosseguimento dos estudos;
- Preparar para o mercado de trabalho e vida de cidadão;
- Aprimorar a formação humana abordando tópicos como ética e pensamento crítico; e
- Compreensão do ambiente científico e as diferenças entre teoria e prática.

O sistema brasileiro conta também com o ensino superior, técnico e demais ensinos especiais. Estas categorias de ensino não tem caráter obrigatório ao cidadão brasileiro.

2.2 Setores da Educação Frente à Educação Privada

A educação no Brasil é fornecida tanto pelo governo, seja ele municipal, estadual ou federal, quanto por instituições privadas. A educação privada é fundamental para a formação do brasileiro, visto que ocupa uma parcela relevante do corpo estudantil do país. De acordo com o Censo Educacional de 2019 a federação possui o perfil educacional conforme demonstrado na Figura 2.1 (MEC, 2019):

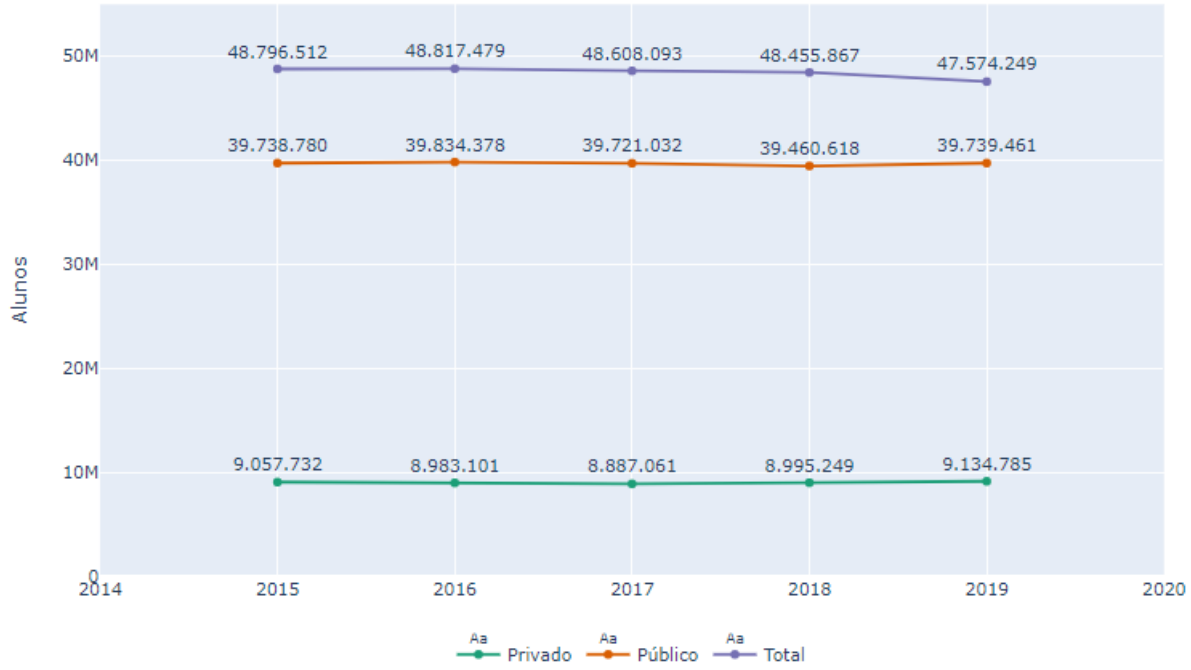

Figura 2.1: Volume de alunos no sistema de Educação Básica brasileiro.

Fonte: Adaptado de MEC (2019)

Segundo o Censo Educacional de 2019 (MEC, 2019), a educação pública conta com 38 milhões de alunos matriculados no ensino básico. A educação privada, por sua vez, conta com cerca de 9 milhões de alunos. Em termos percentuais, conforme ilustrado na Figura 2.2 pode-se afirmar que a educação privada consiste de 32% dos alunos no país.

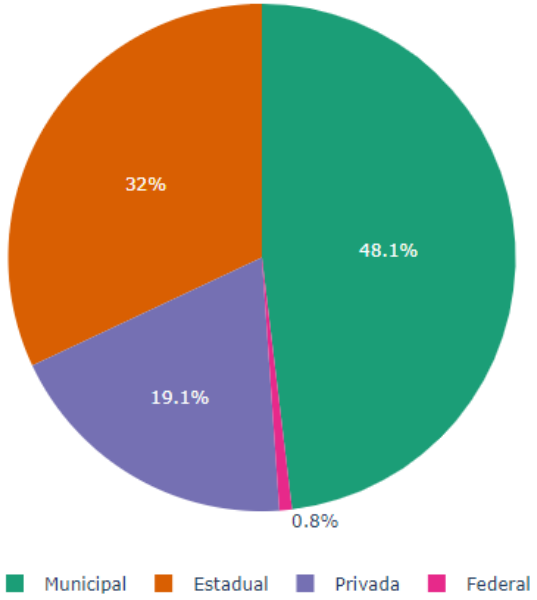

Figura 2.2: Distribuição dos alunos brasileiros na Educação Básica por órgão responsável.

Fonte: Adaptado de MEC (2019)

Dado tal retrato, é possível afirmar que existe um relevante impacto econômico no país devido à educação privada, um setor que é responsável parte da educação dos brasileiros.

2.3 Importância Social e Impacto da Evasão

A evasão estudantil é um tema de elevada relevância na área da educação. Tendo direto impacto no futuros dos estudantes, é de interesse dos responsáveis pelo ensino dos alunos, sejam estes nações, empresas, profissionais da educação ou família, a amenização deste problema (MEC, 2019; BRASIL, 2021).

Embora não exista um completo consenso na literatura sobre a definição de evasão escolar (FILHO; ARAÚJO, 2017), diversos autores (CERATTI, 2008; BATISTA; SOUZA; OLIVEIRA, 2009; FILHO; ARAÚJO, 2017; SOUSA et al., 2018) definem evasão como o abandono dos estudos durante o ano escolar. Sendo assim, *evasores*, isto é, alunos que abandonaram os estudos durante o ano letivo, são o ponto focal de trabalhos sobre este tópico.

De acordo com um estudo realizado pela Brasil (2021), a evasão em conjunto da reprovação e da disparidade idade-série, são comuns consequências do fracasso escolar. No ano de 2019 cerca de 600 mil estudantes abandonaram seus estudos (BRASIL, 2021).

De acordo com uma pesquisa (KAYONDA et al., 2021) realizada sobre famílias onde existe abandono escolar notam como consequências menor capacidade econômica, menor reconhecimento social, envolvimento com atividades ilícitas e problemas de auto-estima.

Além de piores condições na vida adulta, a evasão estudantil em adolescentes infratores tem correlação com uma maior taxa de reincidência de detenção (NA, 2017), embora não necessariamente implique ao cometimento de novas infrações.

Além dos impactos sociais, a evasão tem grande impacto econômico. Em um estudo realizado sobre ex-estudantes latino-americanos (ADELMAN; SZEKELY, 2016), foi encontrado que a taxa de desemprego para indivíduos que não completaram seus estudos é maior quando comparada com os que finalizaram o Ensino Médio. Além disso, é notado também que os salários de indivíduos que completaram seus estudos até o Ensino Médio são maiores dos que suas contrapartes evasoras, com essas disparidade aumentando em conjunto com o nível de educação dos indivíduos.

2.4 Problemas nos Estudos e Evasão

A educação é um fator importante no futuro de um cidadão. Entretanto existem diversas dificuldades e problemas associadas ao processo educacional e as vidas dos alunos. A reprovação escolar é um fator que incentiva a evasão dos estudos (LEON; MENEZES-FILHO, 2002), aparecendo em até 21% dos casos de evasão escolar no decorrer dos anos. A nota do aluno também possui correlação com a desistência nos estudos, onde estudantes com menores notas tendem a desistir mais (BARDACH et al., 2019; BRASIL, 2021).

É comum também observar uma maior evasão em casos onde a renda familiar é menor, ou caiu. Estudos com base nos dados históricos do Censo Escolar (MEC, 2019) afirmam que há uma correlação negativa entre renda e evasão. Isto é, um estudante mais rico, ou mais estável financeiramente, é menos provável a evadir os estudos (LEON; MENEZES-FILHO, 2002). Os fatores financeiros que influenciam na evasão estudantil pela existência de populações carentes e as dificuldades que essas populações passam (LEON; MENEZES-FILHO, 2002).

No âmbito internacional, por exemplo, é frequente a evasão nas escolas devido à problemas financeiros (como por exemplo uma baixa renda), de desempenho (notas baixas ou desempenho insatisfatório) ou até contextuais (sazonais ou dependentes de ambientes à qual o aluno frequenta). Estudos realizados com estudantes do ensino superior (BARDACH et al., 2019) constatam tais fatos. Entretanto, em países mais desenvolvidos, existem políticas que visam amenizar tais incidentes, como por exemplo o fornecimento de bolsas aos estudantes (MIWA et al., 2015) e maior participação na vida do estudante por parte da escola (BARDACH et al., 2019) nas suas metas pessoais.

2.5 *Churn* e Evasão Estudantil

Em mercados que oferecem serviços como a saúde, educação ou estética uma relevante parcela dos clientes são aqueles que já fizeram uso do serviço previamente. Sendo assim, uma não negligenciável parte da renda da empresa consiste da retenção da clientela. Por sua vez, devem manter a qualidade do serviço para aumentar a fidelização (LEJEUNE, 2001). Sendo assim, se faz necessário estudar motivos para retenção de clientes e os motivos para sua saída.

2.5.1 *Churn*

O conceito de *churn* pode ser definido como "a rotatividade anual do mercado de base" (STROUSE, 1999). Pode ser descrito como a razão entre os clientes que pararam de aderir a um serviço com relação ao total de clientes atingidos.

Churn também pode ser definido como o processo em que pagantes de um serviço contínuo, seja ele pago *a priori* ou *a posteriori*, trocam para um outro serviço da concorrência (BERSON; THEARLING, 1999).

De tal maneira, a previsão de *churn* pode ser definida de diferentes formas. Uma maneira de definir previsão de *churn* é a probabilidade de um cliente mudar de serviço (DAHIYA; BHATIA, 2015).

No contexto da gestão do relacionamento com o cliente, o *churn* é uma métrica que mede a saúde do negócio e o prospecto de lucros futuros. Além do *churn*, ou retenção de cliente, existem medidas como aquisição, extensão e seleção de clientes que também são importantes para medir a saúde do negócio (KOMENAR, 1997).

2.5.2 Previsão de *Churn*

A previsão de *churn*, devido a sua importância no mundo dos negócios, é um tópico extensivamente abordado na literatura em diferentes áreas do conhecimento (AHN et al., 2020). A área da telecomunicação, seja telefonia móvel, televisão à cabo, provedores de internet ou serviços similares é uma área onde este tipo de métrica é constantemente avaliada e estudada (KAMALRAJ; MALATHI, 2013; ALMANA; AKSOY; ALZAHRANI, 2014; BANDARA; PERERA; ALAHAKOON, 2013).

Com o passar dos anos, as técnicas para previsão de *churn* tem evoluido constantemente. As técnicas usadas para atacar o problema de previsão de *churn* são apresentadas pela Figura 2.3, adaptada de Ahmed et al. (2017).

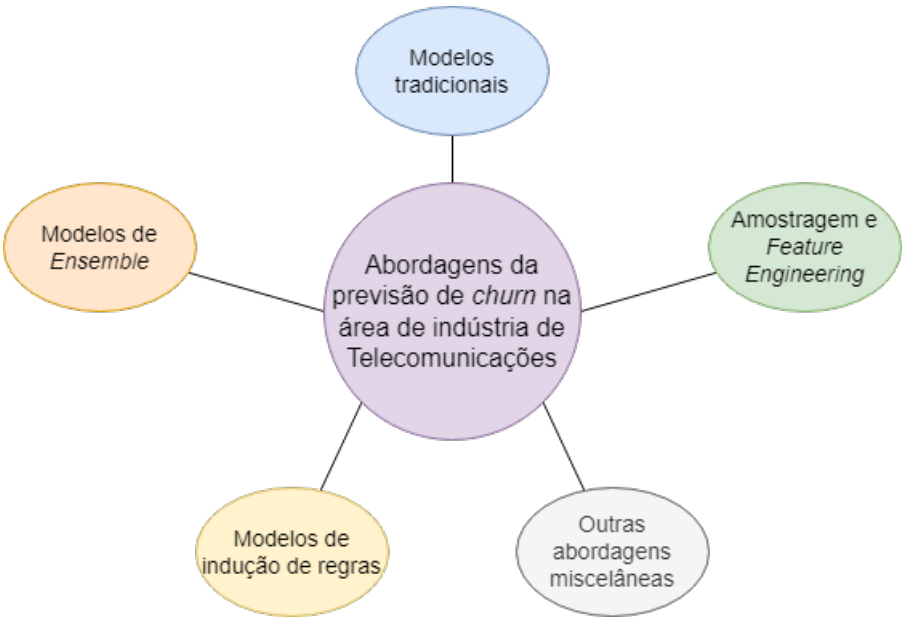

Figura 2.3: Métodos para previsão de *churn*.

Fonte: Adaptado de Ahmed et al. (2017)

Uma maneira clássica de abordar problemas de *churn* em diferentes áreas é por meio do uso de algoritmos classificadores tradicionais, como *Naive Bayes*, Máquinas de vetores de suporte (*Support Vector Machines*, SVM), Redes Neurais e Árvores de decisão (AHMED et al., 2017). Tais abordagens são extensivamente abordadas na literatura de aprendizagem de máquina (AHMED et al., 2017; HUANG; KECHADI; BUCKLEY, 2012) e tornaram-se tarefas triviais frente à outros desafios apresentados à profissionais que trabalham com tais ferramentas.

Esses algoritmos tentam, a partir do conjunto de dados, gerar um único classificador, também chamado de classificador monolítico, que classifique as entradas. Tais classificadores já apresentaram resultados satisfatórios em casos na telecomunicação (QI et al., 2006; HUANG; KECHADI; BUCKLEY, 2012; VILLARREAL et al., 2020; DUMITRACHE; NASTU; STANCU, 2020), servindo como motor para sistemas de aviso prévio.

Aliado ao uso de classificadores, sejam monolíticos ou não, pode-se utilizar de técnicas de re-amostragem e engenharia de atributos (*feature engineering*) para melhorar os resultados (LYKOURENTZOU et al., 2009).

2.5.3 Evasão Estudantil

Devido à relevância da evasão estudantil, este tópico é abordado do ponto de vista de educação, onde é usada como métrica para avaliação de qualidade de ensino (LAU, 2003;

HABLEY; MCCLANAHAN, 2004), da psicologia, onde é visto os impactos na vida do estudante (O'KEEFFE, 2013) e da computação, onde explora-se o problema com meio de algoritmos (MÁRQUEZ et al., 2016).

Devido à maneira que o mercado de escolas privadas se apresenta, isto é, por meio mensalidade recorrente em troca da educação uma evasão estudantil, configura um problema de *churn* na área da educação (QUARTI; FIGINI; GIUDICI, 2009).

Embora apresente semelhanças com o problema de *churn* em demais mercados, os motivos que levam à desistência do estudante durante seus estudos são diversos. Um estudante pode desistir dos estudos, por exemplo, por falta de engajamento, dificuldades financeiras e pouco incentivo acadêmico (BEAL; NOEL, 1980). Além disso, a competição nesse tipo de mercado tem impacto direto (MITCHELL; COLES, 2004) na taxa de retenção estudantil.

Entretanto, a implementação de medidas para a resolução de tais problemas é difícil e lenta, além de evoluir conforme o tempo (TINTO, 2006). De tal maneira, é um problema que requer constante atenção.

2.6 Classificadores e modelos interpretáveis

A existência de situações onde deseja-se ter uma estimativa de valores ou categorias de casos futuros é constante em diversas áreas. Dentro do contexto de aprendizagem de máquina, o uso de base de dados para alimentar algoritmos para abordar esse problema é constante (QUINLAN, 2014; BERKSON, 1944; BREIMAN, 2001; FREUND; SCHAPIRE, 1997; CHEN; GUESTRIN, 2016). Algoritmos, ou modelos, usados para gerar rótulos ou categorias para exemplos apresentados podem ser chamados de classificadores (RUSSELL; NORVIG, 2010).

Classificadores são uma ferramenta poderosa dentro da Inteligência Artificial, tendo diversos casos de aplicações em problemas fora do ambiente acadêmico (AHMED et al., 2017; BANDARA; PERERA; ALAHAKOON, 2013; AHN et al., 2020; MÁRQUEZ et al., 2016; LYKOURENTZOU et al., 2009; SALES; BALBY; CAJUEIRO, 2016). Embora a adoção de classificadores auxilie na abordagem de uma miríade de problemas, o processo de seleção do rótulo por parte do algoritmo não é necessariamente óbvio para seus usuários. De tal maneira, é possível categorizar algoritmos com relação a sua interpretabilidade.

2.6.1 Modelos caixa preta e caixa branca

Devido a maior adesão de modelos para atender problemas e as necessidades apresentadas pelas áreas que os aplicam, o debate entre abordagens caixa preta e caixa branca se torna cada vez mais relevante.

Um modelo caixa preta é um modelo que tenta criar um modelo matemático complexo para resolver o problema em questão. Exemplos de modelos matemáticos complexos por trás de modelos caixa preta são uma série de equações complexas (como as redes neurais e SVM) ou necessidade do entendimento de funções de distância e espaço de representação (como KNN) (LOYOLA-GONZALEZ, 2019). A Figura 2.4, adaptada de (FUNG et al., 2021), trás como exemplo algoritmos de caixa preta, com difícil compreensão de seu resultado.

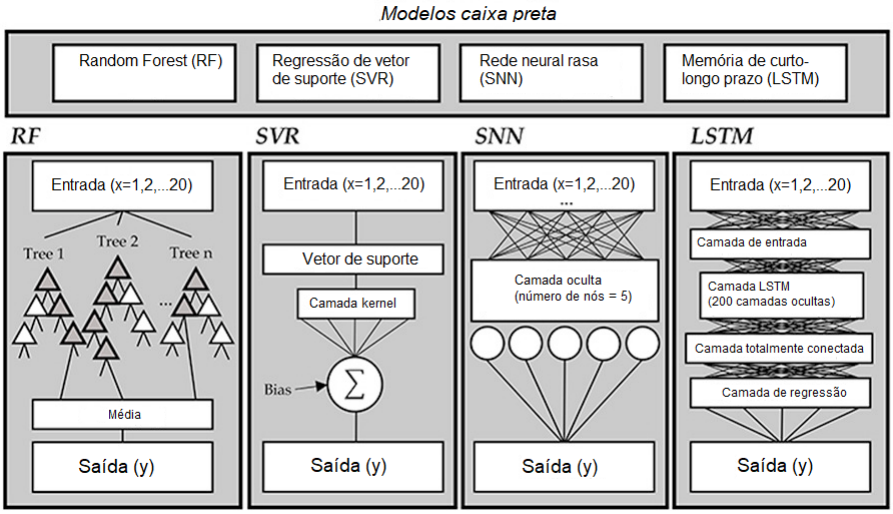

Figura 2.4: Exemplos de algoritmos caixa preta.

Fonte: Adaptado de Fung et al. (2021)

Por sua vez, um modelo caixa branca tem seu resultado facilmente explicado. Modelos caixa branca fazem uso de padrões e regras para chegar à sua saída que por sua vez são mais fáceis de serem interpretados. Árvores de decisão, por exemplo, são facilmente entendidas e servem como exemplo de modelo caixa branca (LOYOLA-GONZALEZ, 2019). A Figura 2.5, adaptada de (FUNG et al., 2021), trás como exemplo algoritmos de caixa branca comumente usados em problemas de regressão.

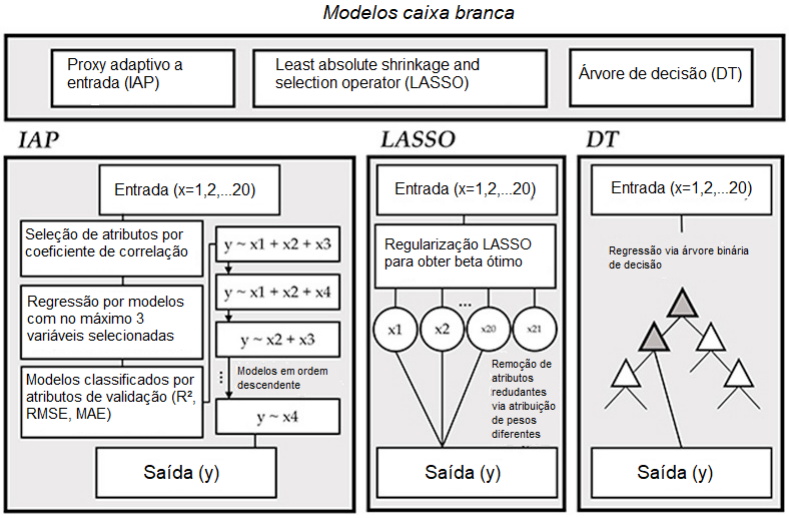

Figura 2.5: Exemplos de algoritmos caixa branca.

Fonte: Adaptado de Fung et al. (2021)

Em indústrias e mercados onde há grande necessidade de entender o motivo por trás de uma classificação, além de sua acurácia, há uma tendência a se utilizar modelos de caixa branca. A lei americana de Oportunidade de Crédito Equalitária (USA, 1974), por exemplo, proíbe distinção por meio de alguns critérios, como sexo e cor, na hora de prover crédito. Sendo assim, se faz necessário obter um resultado que provenha a explicação para sua saída, garantindo a não utilização de informações proibidas em sistemas preditivos.

2.6.2 Técnicas para Avaliação e Explicação de Resultados de Modelos

Levando em consideração que nem sempre modelos de caixa branca fornecem os melhores resultados para todos os problemas, a utilização de técnicas para interpretação de modelos de classificação é de grande utilidade.

Com o passar dos anos múltiplas técnicas foram utilizadas para realizar a interpretação, ou explicação, de modelos de aprendizado de máquina (RIBEIRO; SINGH; GUESTRIN, 2016; ŠTRUMBELJ; KONONENKO, 2014; SHRIKUMAR; GREENSIDE; KUNDAJE, 2017; LUNDBERG et al., 2018). Dentre elas, algumas técnicas tem ganhado muito destaque devido ao seu sucesso em prover satisfatórias explicações para algoritmos complexos, como redes neurais (BACH et al., 2015; LUNDBERG; LEE, 2017).

2.6.3 Classificadores

Dentro da área de aprendizado de máquina existem diversos algoritmos próprios para tarefas de classificação (BERKSON, 1944; QUINLAN, 2014; BREIMAN, 2001; CHEN;

GUESTRIN, 2016). Devido a diferenças na maneira que operam, certos algoritmos tendem a oferecer melhores resultados em alguns problemas quando comparados a outros. A escolha dos algoritmos a serem avaliados é um importante passo em tarefas de classificação, onde leva-se em consideração métricas, interpretabilidade, desempenho e manutenibilidade na hora da escolha final do modelo (RUSSELL; NORVIG, 2010).

Um classificador é dito monolítico quando apenas uma instância de algoritmo classificador é levada em consideração para a geração do rótulo final. Entretanto, é possível combinar o resultado de diversas instâncias de um ou mais algoritmos a fim de obter resultados melhores. A adoção de múltiplos algoritmos é chamada de *ensemble* e é uma opção que aproveita as qualidades de diversas instâncias de algoritmos (RUSSELL; NORVIG, 2010).

2.6.4 Regressão Logística

O modelo de regressão logística, ou modelo *logit* (BERKSON, 1944), é um algoritmo classificador para casos onde a variável alvo é binária, isto é, assume apenas dois valores. O modelo assume a independência dos atributos no conjunto de dados para determinar se uma instância é positiva ou negativa.

Por meio do cálculo dos logaritmos das chances dos atributos assumirem os respectivos valores, a regressão logística calcula a probabilidade da instância assumir cada valor. Dadas as probabilidades isoladas, é possível determinar qual a classe com maior probabilidade.

O modelo é dito caixa-branca, visto que fornece os impactos de cada atributo na classificação final, e portanto, viabilizando a interpretação dos resultados apresentados.

2.6.4.1 Árvore de decisão

A Árvore de Decisão, ou *Decision Tree*, é um algoritmo classificador baseado em Árvores de Busca. Tendo como uma de suas mais famosas implementações o C4.5 (QUINLAN, 2014), a Árvore de Decisão consiste na divisão sucessiva da base de dados. O critério para divisão da base adotado para a divisão almeja em selecionar os atributos que melhor dividem a base de dados tendo em mente a classe alvo.

Por meio do critério adotado no algoritmo, obtém-se os atributos que tem maior impacto na classe final. Gerando assim uma sequência lógica de comparações para a definição da saída do algoritmo. Critérios comumente adotados para Árvores de decisão são o grau de impuridade GINI e o grau de entropia (QUINLAN, 2014). Esse método

gera então, a partir da base de treino, um conjunto de decisões a serem tomadas ao ser apresentada um exemplo para classificação. Embora eficaz, devido a essas características a falta de dados, a existência de *outliers* ou a inserção parâmetros não-ideais pode levar a Árvore de Decisão a decorarem o conjunto de dados, termo denominado *overfitting*.

A Árvore de Decisão é um modelo caixa branca, e portanto, é capaz de gerar uma interpretação e visualização nativa própria algoritmo para as classificações realizadas. A Figura 2.6 demonstra uma representação obtida por Pedregosa et al. (2011) ao treinar uma Árvore de Decisão, um algoritmo caixa branca.

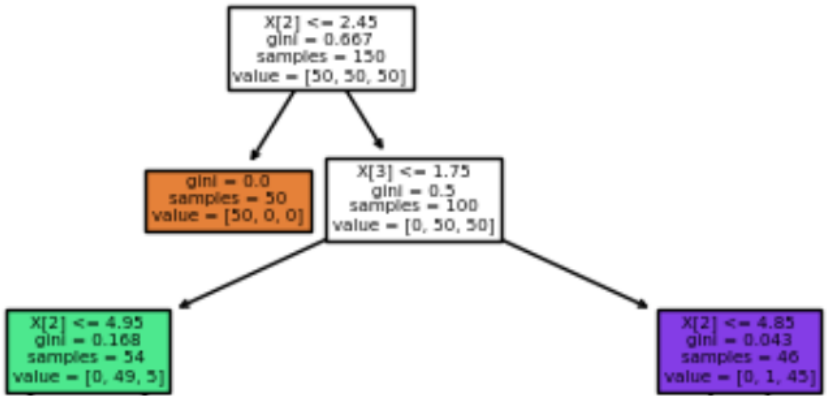

Figura 2.6: Representação gráfica de uma árvore de decisão.

Fonte: Adaptado de Pedregosa et al. (2011)

No exemplo apresentado, pode-se verificar que por exemplo, os atributos $X[2]$ e $X[3]$ tem alto impacto na classe pela qual o classificador dará como saída representada pelo tom e intensidade da cor (PEDREGOSA et al., 2011).

A árvore de decisão é um modelo simples porém seu custo computacional e capacidade de customização fazem com que seja adotada em diversos algoritmos de *ensemble* mais complexos como o *Random Forest* (BREIMAN, 2001) e o *XGBoost* (CHEN; GUESTRIN, 2016).

2.6.4.2 *Random Forest*

A criação de *ensembles* é uma técnica que troca simplicidade dos modelos treinados por uma capacidade de classificar exemplos e problemais mais complexos. Fazendo a combinação de múltiplas Árvores de Decisão, o algoritmo *Random Forest* (BREIMAN, 2001) foi criado com esse objetivo em mente.

Em casos onde as características apresentadas pela Árvore de Decisão são desejáveis, mas o problema a ser estudado possui alta complexidade ou tende a gerar modelos

com *overfit*, o algoritmo *Random Forest* pode apresentar resultados mais próximos dos desejados.

Por meio do treino de múltiplas Árvores de Decisão, as quais recebem um subconjunto dos atributos disponíveis na base, o algoritmo determina a classe dos exemplos apresentados por meio de um sistema de voto simples (BREIMAN, 2001). Sendo assim, em casos de classificação a classe apresentada como a mais provável pela maioria dos modelos é dada como resposta final do algoritmo.

Devido ao fato que as *Random Forests* possam conter uma multitude de árvores diferentes, as quais podem apontar diferentes resultados, o algoritmo perde a interpretabilidade das Árvores de Decisão. Sendo assim, um modelo caixa preta. Entretanto, em casos onde a interpretabilidade do modelo não é prioritária, o algoritmo é uma poderosa ferramenta em problemas de classificação.

2.6.4.3 *AdaBoost*

Algoritmos de *ensemble* com modelos criados de maneira aleatória podem apresentar bons resultados, como no caso das *Random Forests*. Entretanto, em casos onde a base de dados apresenta entradas com maior complexidade quando comparados aos demais exemplos, técnicas como o *Adaptive Boosting*, ou *AdaBoost*, podem ser aplicadas (FREUND; SCHAPIRE, 1997).

O algoritmo *AdaBoost* implementa o processo de *boosting*. O *boosting* consiste de múltiplas rodadas de treino, onde novos modelos são treinados e adicionados ao *ensemble* com intuito de melhor interpretar o problema em questão. Esse processo é repetido até o erro estar dentro da margem configurada.

O *AdaBoost* realiza diversas rodadas de treino com instâncias da base de dados original. Fazendo uso da atribuição de pesos aos exemplos com maior taxa de erro durante cada treino, o algoritmo inclui os exemplos de maior complexidade em futuras rodadas mais frequentemente. De tal maneira, o algoritmo se adapta a cada iteração, potencialmente tornando o modelo mais sensível aos diferentes tipos de instância (FREUND; SCHAPIRE, 1997).

O *AdaBoost* é um algoritmo de *ensemble*, que normalmente adota classificadores base simples como a Árvores de Decisão (FREUND; SCHAPIRE, 1997), mas implementações do algoritmo podem usar outros classificadores base (PEDREGOSA et al., 2011). Devido a grande quantidade de modelos que podem ser gerados durante o processo de *boosting*, o algoritmo *AdaBoost* é dito um algoritmo caixa preta.

2.6.4.4 *XGBoost*

De maneira similar ao *AdaBoost*, o *XGBoost*, ou *eXtreme Gradient Boosting*, é um algoritmo de *ensemble* que aplica o processo de *boosting* para melhor se adaptar ao conjunto de treino. Entretanto, o *XGBoost* apresenta diferenças significativas no seu funcionamento (CHEN; GUESTRIN, 2016).

Enquanto o *AdaBoost* usa das entradas com maior erro durante cada iteração para o treino de novos modelos no *ensemble*, o *XGBoost* define uma função de perda. Função esta que mede o erro dos modelos treinados até uma iteração com o resultado esperado (CHEN; GUESTRIN, 2016).

Durante cada iteração, a perda é avaliada e um novo modelo, melhor que os anteriores, é adicionado ao *ensemble*. Esta característica tenta apresentar resultados corretos desde o início do processo, em contrapartida ao *AdaBoost*, que os corrige durante o processo (AMAZON, 2022).

O algoritmo *XGBoost* possui implementação otimizada para rodar em múltiplos núcleos de CPU de maneira paralela, fazendo o processo de treino de maneira mais rápida (CHEN; GUESTRIN, 2016; AMAZON, 2022). O algoritmo é dito caixa preta, por motivos similares ao *AdaBoost*.

2.7 Técnicas de Interpretação de Modelos

Décadas atrás, a comunidade científica se concentrava em propor modelos para problemas teóricos, entretanto, atualmente a adesão de técnicas computacionais para resolver problemas no mundo real é crescente (LOYOLA-GONZALEZ, 2019). Sendo assim, existem diversos esforços na área de aprendizagem de máquina para gerar modelos os quais não apenas gerassem resultados mas também ajudassem a interpretar o problema em questão.

2.7.1 Local Interpretable Model-agnostic Explanations (LIME)

O LIME, ou *Local Interpretable Model-agnostic Explanations*, é um modelo para gerar interpretações à partir de modelos sem a necessidade de ser um modelo caixa branca (RIBEIRO; SINGH; GUESTRIN, 2016). Isto é, conforme o nome afirma, um gerador de explicações agnóstico à modelos.

O LIME tem como objetivo não apenas gerar uma explicação de quais atributos geraram o resultado, e sim uma representação do problema que evidencia a partipação destes atributos (RIBEIRO; SINGH; GUESTRIN, 2016). Em um problema de classifi-

cação, por exemplo, o LIME pode gerar as saídas apresentadas na Figura 2.7 para uma imagem classificada como Labrador, Violão e Guitarra (RIBEIRO; SINGH; GUESTRIN, 2016), a qual tem como atributos os próprios *pixels* da imagem.

Figura 2.7: Saída obtida pelo LIME em um algoritmo que classificou a imagem como Labrador (21%), Violão (24%) e Guitarra (32%).

Fonte: Adaptada de Ribeiro, Singh e Guestrin (2016)

O LIME parte da idéia de treinar modelos interpretáveis, modelos caixa branca como árvores de decisão e regressões lineares, a partir dos resultados de um modelo não-interpretável, como modelos caixa preta como o XGBoost (CHEN; GUESTRIN, 2016).

De acordo com Molnar (2020) o algoritmo LIME (RIBEIRO; SINGH; GUESTRIN, 2016) pode ser explicado de maneira simples com o seguinte algoritmo:

- Seleciona-se uma instância de interesse x a partir da qual deseja-se a interpretação do modelo de caixa preta f;
- Pertuba-se o conjunto de dados, adulterando valores e obtém-se as previsões do modelo de caixa preta a partir desse novo conjunto de pontos;
- Atribui-se pesos aos novos pontos, de acordo com a proximidade à instância de interesse;
- Treina-se um modelo interpretável g ponderado no conjunto de dados com as novas variações; e
- Explica-se a instância de interesse por meio da interpretação do novo modelo interpretável.

De tal maneira, obtém-se uma aproximação da resposta do modelo sob tais condições. A Figura 2.8 (RIBEIRO; SINGH; GUESTRIN, 2016) ilustra essa idéia.

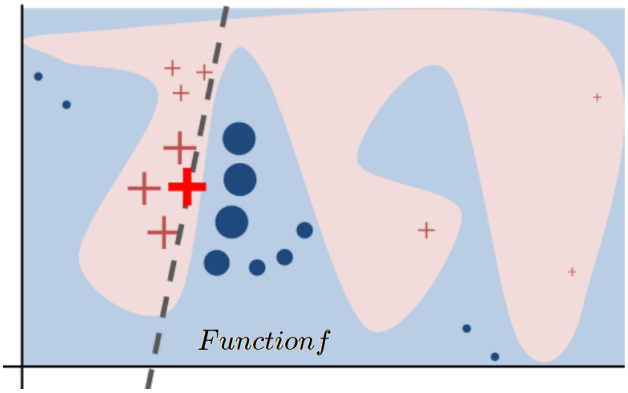

Figura 2.8: Exemplo de funcionamento do LIME.

Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016)

A Figura 2.8 representa f por meio da área azul. Os demais pontos gerados pelo algoritmo, variações de x, tem semelhança à entrada dada pelo tamanho e são representados por cruzes e círculos.

O LIME pode ser aplicado em imagens, texto ou dados tabulares, conforme demonstram Ribeiro, Singh e Guestrin (2016) em um conjunto de dados sobre cogumelos venenosos e comestíveis, usado em tarefas de classificação (LINCOFF; SCHLIMMER, 1981). O conjunto Mushroom (LINCOFF; SCHLIMMER, 1981) conta com dados respectivos à odor, tipo de esporos e demais informações dos cogumelos. A Figura 2.9 demonstra uma saída gerada pelo LIME na classificação de um cogumelo do conjunto de dados.

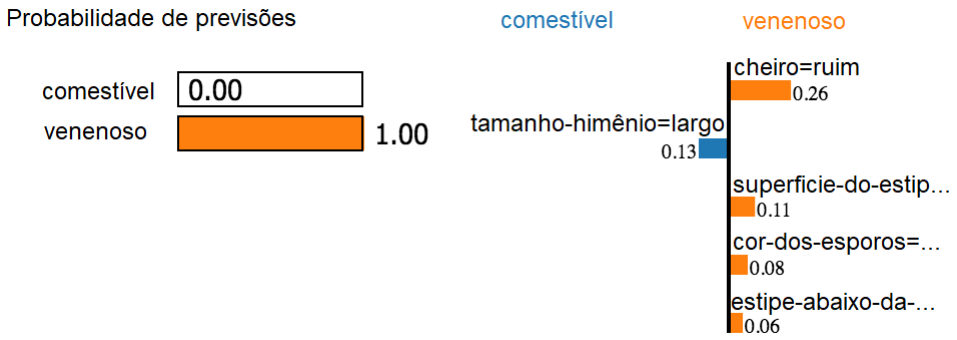

Figura 2.9: Exemplo do LIME aplicado ao conjunto de dados *Mushroom*.

Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016), Lincoff e Schlimmer (1981)

Conforme demonstrado na Figura 2.9, o classificador indicaria que o cogumelo testado seria um cogumelo venenoso. Isso se deve pelo cheiro pútrido, cor dos esporos e a textura do cogumelo (RIBEIRO; SINGH; GUESTRIN, 2016). Um fator que contribuiria para ele ser comestível, segundo o LIME, é o tamanho dos himênios, o qual não foi suficiente para o classificador atribuir a classe de comestível ao cogumelo (RIBEIRO;

SINGH; GUESTRIN, 2016).

2.7.2 Valores de Shapley

Outra maneira para gerar explicações para modelos é usar conceitos de teoria dos jogos. Afirmando de maneira análoga que cada instância no conjunto de dados é um jogo, o ganho é a previsão final do modelo e os atributos são jogadores, calcular os pagamentos individuais de cada jogador de maneira justa é medir a contribuição de cada jogador no resultado final do jogo (SHAPLEY, 1951).

De tal maneira, calcular a contribuição de um atributo na classificação de uma instância em um conjunto de dados é ver a sua contribuição no resultado final de um modelo.

Em modelos de regressão linear, por exemplo, as contribuições de cada atributo são dadas pelo peso do atributo e o seu valor na instância avaliada. Entretanto, em modelos mais complexos esse valor de contribuição é mais difícil de ser calculado. O LIME (RIBEIRO; SINGH; GUESTRIN, 2016) usa de modelos e perturbações no conjunto de dados para a geração de pesos correspondentes à cada atributo para então gerar sua contribuição.

Por sua vez, os valores de Shapley são outra maneira de atribuir as contribuições a cada atributo, tratando o problema como uma maneira de distribuir justamente o "pagamento" entre diferentes conjuntos de "jogadores". Um valor de Shapley é a contribuição marginal média de um atributo entre todas as possíveis permutações no conjunto de dados.

Por exemplo, assume-se o conjunto de dados, dado na Tabela 2.1, que contém dados dos ganhos em um jogo hipotético que pode ser jogador sozinho ou em dupla:

Tabela 2.1: Dados fictícios para o valor do ganho em um jogo

j_1 **- Jogador 1 jogou**	j_2 **- Jogador 2 jogou**	val **- Ganho no jogo**
0	0	0
1	0	1
0	1	2
1	1	4

Deseja-se por exemplo obter a contribuição do atributo j_1, em um modelo ideal para a previsão do ganho G, treinado no conjunto de dados descrito pela Tabela 2.1. A contribuição pode ser estimada a partir do cálculo do valor de Shapley do atributo. O valor de Shapley de um atributo j é dado pela Equação 2.1 (SHAPLEY, 1951):

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right) \tag{2.1}$$

Em outras palavras, para calcular o valor de Shapley ϕ de um do atributo j, avalia-se a diferença no resultado val do modelo quando o atributo é adicionado em cada uma das possíveis combinações S dos p atributos possíveis do conjunto de dados.

De tal maneira, encontra-se que o valor de Shapley para a contribuição do atributo j_1 no exemplo da Tabela 2.1 pode ser dado, de maneira simplificada, por:

$$\phi_1(val) = \frac{val(1) + (val(1,2) - val(2))}{2}$$

A partir da Tabela 2.1 têm-se que os seguintes valores são verdade quando o valor val do modelo é avaliado. Valores estes que prevem o ganho no jogo:

- Ganho val quando apenas o Jogador 1 joga assume valor: $val_1 = 1$;
- Ganho val quando apenas o Jogador 2 joga assume valor: $val_2 = 2$; e
- Ganho val quando ambos Jogador 1 e 2 jogam assume valor: $val_{1,2} = 4$.

Portando, o valor de Shapley ϕ do atributo j_1 é calculado por:

$$\begin{aligned}
\phi_1(val) &= \frac{val(1) + (val(1,2) - val(2))}{2} \\
\phi_1(val) &= \frac{1 + (4 - 2)}{2} \\
\phi_1(val) &= 1.5
\end{aligned}$$

Por sua vez, o valor de Shapley do atributo pode ser dado por:

$$\begin{aligned}
\phi_2(val) &= \frac{val(2) + (val(1,2) - val(1))}{2} \\
\phi_2(val) &= \frac{2 + (4 - 1)}{2} \\
\phi_2(val) &= 2.5
\end{aligned}$$

Finalmente, em um jogo justo, a divisão do ganho total, o qual assume valor 4, entre os jogadores j_1 e j_2 seria de 1.5 e 2.5, respectivamente. Analogamente, a influência da presença do atributo j_1 na previsão do resultado do modelo val é de 1.5, enquanto a influência de j_2 é de 2.5. Os valores obtidos não necessariamente correspondem aos reais

valores levado em consideração pelo modelo, mas servem como uma estimativa na hora de explicar o resultado obtido.

Os valores de Shapley são uma poderosa ferramenta para determinar as contribuições dos atributos, e portanto, ajudar na interpretação de um modelo. Entretanto, a necessidade de testar todos as possíveis combinações para obter um resultado que não está necessariamente correto é um ponto negativo do método.

2.7.3 SHAP

Usado na teoria dos jogos e em aprendizado de máquina, os valores de Shapley são números reais que podem ser interpretados como o impacto do valor de um atributo no resultado final da classificação, assumindo um ponto base inicial (LUNDBERG; LEE, 2017). Entretanto, os valores de Shapley são altamente custosos de serem calculados pela sua definição original (SHAPLEY, 1951) e tentativas de estimar esses valores são mais viáveis computacionalmente.

O SHAP, ou *SHapley Additive exPlanations*, (LUNDBERG; LEE, 2017) é um método e biblioteca para geração de intepretações para modelos que unifica diversos algoritmos e estimativas para cálculos de valores de Shapley, além de incluir gráficos para melhor interpretação dos modelos avaliados. O método foi construído com o intuito de ser uma forma mais robusta e completa para descrever as participações dos atributos no resultado final de um modelo preditivo.

Para determinar a importância de cada atributo da entrada a ser classificada, o SHAP usa dos valores obtidos por diferentes técnicas, como o LIME (RIBEIRO; SINGH; GUESTRIN, 2016) ou o Método de atribuições aditivas de características (ŠTRUMBELJ; KONONENKO, 2014) e os trata como valores de Shapley (LUNDBERG; LEE, 2017) para alimentar o núcleo do algoritmo. Tal abordagem trás como vantagem o menor custo computacional quando comparado ao cálculo dos valores de Shapley, que aumentam exponencialmente conforme o número de atributos cresce (MOLNAR, 2020).

O SHAP conta com diversos núcleos, cada qual sendo mais adequado para diferentes tipos de problemas. Alguns dos métodos oferecidos pelo SHAP são:

- *Kernel*: utiliza o LIME (RIBEIRO; SINGH; GUESTRIN, 2016) para a geração dos pesos dos atributos nos valores de Shapley;
- *Gradient*: utiliza de conceitos como gradientes esperados e tem como vantagem levar o conjunto de dados completo como base de cálculos para os valores de Shapley;
- Linear: calcula analiticamente os valores de Shapley reais; e

- Profundo: utiliza uma versão do DeepLift (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) para estimar os valores de pesos dos valores de Shapley, ideal para modelos de aprendizagem profunda.

Cada núcleo do SHAP parte de uma série de hipóteses sobre o conjunto de dados, as quais devem ser testadas antes de aplicar o algoritmo para gerar a saída do modelo (LUNDBERG; LEE, 2017).

O SHAP tem encontrado uso em diversas tarefas. A Figura 2.10 (LUNDBERG; LEE, 2017) demonstra a saída do algoritmo após ser aplicado em uma rede neural convolucional.

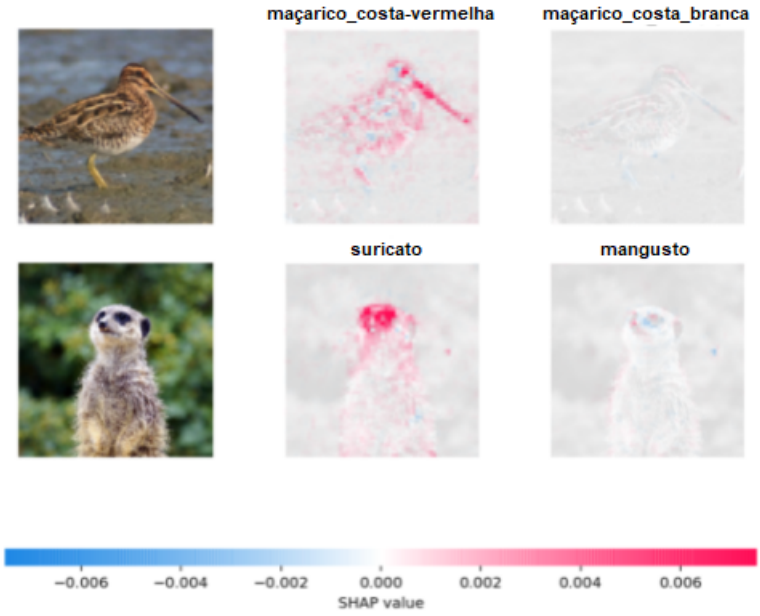

Figura 2.10: Exemplo do SHAP sendo aplicado à imagens alimentadas em uma rede neural convolucional.

Fonte: Adaptada de Lundberg e Lee (2017)

O algoritmo pode ser aplicado também à classificadores mais simples e a casos onde não há representação gráfica de cada exemplo. A figura 2.11 (LUNDBERG; LEE, 2017) demonstra o SHAP sendo aplicado ao conjunto de dados Iris (FISHER, 1936).

O conjunto de dados Iris descreve dados de comprimento e largura das pétalas e sépalas de 150 flores do gênero *Iris* contemplando as espécies *Iris Setosa*, *Iris Versicolour* e *Iris Virginica* (FISHER, 1936). É um conjunto de dados comumente utilizado em tarefas de aprendizado de máquina (LUNDBERG; LEE, 2017).

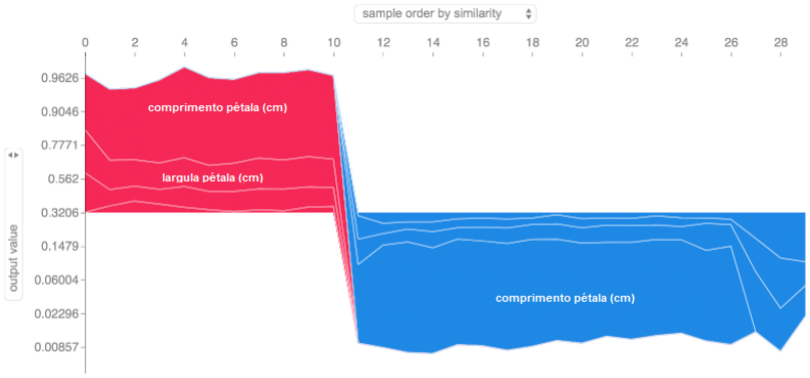

Figura 2.11: Exemplo do SHAP sendo aplicado ao conjunto de dados Iris.

Fonte: Adaptada de Lundberg e Lee (2017), Fisher (1936)

O gráfico apresentado na Figura 2.11 indica os valores de Shapley (SHAPLEY, 1951) calculados para o conjunto de dados Iris (FISHER, 1936) inteiro. Um atributo que tem impacto positivo na saída do modelo SHAP recebe a cor vermelha, caso contrário recebe a cor azul (LUNDBERG; LEE, 2017). Pode-se interpretar a partir do gráfico que, nas instâncias onde o comprimento e largura das pétalas era menor que 10, o modelo tendia a dar uma saída maior no *kernel* do algoritmo SHAP (LUNDBERG; LEE, 2017). Em comprimentos e larguras maiores que 10, o *kernel* do SHAP tinha uma saída menor. O valor de saída do *kernel* do SHAP influencia a classe final atribuída à instância (LUNDBERG; LEE, 2017). O gráfico também apresenta o valor de 0.3206 como o *base value* do modelo, ou o valor esperado pelo SHAP como a saída de um exemplo classificado pelo modelo.

Além de poder interpretar o impacto dos valores dos atributos na saída do algoritmo, pode-se validar também o quão importante o atributo é para tal classificação. Essa importância é dada pela altura da área respectiva ao atributo (LUNDBERG; LEE, 2017). No caso descrito pela Figura 2.11, pode-se notar que o impacto do comprimento da pétala é muito maior que o impacto da sua largura na tarefa de classificação de espécies do conjunto de dados Iris (LUNDBERG; LEE, 2017; FISHER, 1936).

2.8 Considerações Finais

Tendo em mente o direito a educação ao brasileiro e a sua participação durante a infância, adolescência e vida adulta, pode-se ressaltar que é uma área de vital importância social e econômica ao país (BRASIL, 1996).

Devido à gama de conteúdos a serem abordados durante os anos de estudo, se faz necessário dividir o ensino brasileiro em três segmentos: Ensino Infantil, Ensino Funda-

mental e Ensino Médio (BRASIL, 1996). Cada segmento educacional tem como ponto focal jovens brasileiros de diferentes idades e contextos sociais, os quais serão expostos a diferentes conteúdos determinados pelo governo (BRASIL, 1996).

Portanto, levando em consideração a complexidade de abordar todos os momentos da vida acadêmica brasileira e considerando também a importância legal, social e econômica dos estudos, é de suma importância o estudo de motivos que levam a evasão. Diferentes estudos apontam melhores escolas, situação financeira e problemas de contexto como motivadores a saída de um aluno (LAU, 2003; HABLEY; MCCLANAHAN, 2004; MÁRQUEZ et al., 2016).

A escola, vista do ponto de vista de negócio, é um serviço de pagamento recorrente e, portanto, está sujeita a diferentes fenômenos encontrados nesse segmento. Um fenômeno bem documentado é o *churn*, o qual pode ser definido como a troca ou abandono de um serviço. O fenômeno do *churn* é similar à evasão estudantil, visto que o *churn* também se dá devido à características do serviço ofertado (QUARTI; FIGINI; GIUDICI, 2009).

Sendo assim, é possível utilizar de técnicas já comumente adotadas na previsão de *churn*, como modelos preditivos, para prever e atuar em cima de casos onde há possível evasão por parte dos alunos. Existem diferentes abordagens na literatura para a previsão de evasão, na qual algoritmos de aprendizado de máquina como *Support Vector Machines*, árvores de decisão e comitês de classificadores (*ensembles*) demonstraram considerável sucesso (BARDACH et al., 2019; LYKOURENTZOU et al., 2009). Essas técnicas são possíveis devido a massa de dados coletadas com o passar dos anos por parte destes serviços.

Não obstante, saber se um aluno irá evadir não necessariamente é o bastante, visto que não se sabe qual o fator motivante a essa desistência (BARDACH et al., 2019). Além disso, nem todo modelo preditivo apresenta características facilmente explicadas a um time educacional, como é o caso do SVM (CORTES; VAPNIK, 1995; PLATT, 1998). Sendo assim, é de interesse o estudo de técnicas para avaliar os resultados de tais modelos cognitivos.

Dentro da área de aprendizagem de máquina, técnicas como o *Local Interpretable Model-agnostic Explanations* (RIBEIRO; SINGH; GUESTRIN, 2016) e o *SHapley Additive exPlanations* (LUNDBERG; LEE, 2017) apresentaram sucesso na tarefa de melhor compreender os resultados de uma classificação, além de ajudar na formulação de hipóteses por parte de times de negócio.

Levando então os pontos levantados na literatura, um denominador comum entre autores é a gravidade da evasão, ou abandono, escolar. A evasão tem um impacto social, educacional e financeiro na sociedade e esta não pode ser negligenciada. Devido a evasão

ser caso similar ao *churn*, é possível a aplicação de técnicas comuns neste problema. A previsão de *churn* é um tópico largamente abordado e documentado pela literatura. Portanto, devido ao paralelo existente entre características presentes *churn* e evasão, é de interesse verificar o estado da arte para a previsão de evasão.

Capítulo 3

Trabalhos Relacionados

Devido a multidisciplinariedade e a multitude de assuntos relacionados à previsão de evasão, serão avaliados trabalhos correlatos por diferentes ângulos tocantes a esse problema. Considerando a similaridade do *churn* com a evasão, é necessário discutir trabalhos correlatos sobre previsão de *churn*. Sendo assim, a Seção 3.1 aborda trabalhos correlatos na área de previsão de *churn*, apresentando métodos de avaliação dos resultados e algoritmos classificadores utilizados. Em seguida, a Seção 3.2 trata de trabalhos correlatos em explicação de resultados de modelos, com ênfase em *churn*, devido a sua similaridade.

Embora a explicação de resultados de classificadores, até onde se saiba, seja uma área não explorada em problemas de evasão, o problema de previsão de evasão já consta na literatura. De tal maneira, a Seção 3.3 aborda a literatura quanto a fatores que levam à evasão e atributos utilizados em bases de dados para tal problema. Por fim, a Seção 3.4 aborda trabalhos correlatos sob a ótica de previsão de evasão.

3.1 Previsão de *churn*

Para melhor entender o estado da arte e as técnicas utilizadas na previsão de *churn*, optou-se por primeiramente verificar qual o panorama geral da área para em seguida adentrar em exemplos de casos mais específicos.

3.1.1 Panorama geral de *churn*

Diversos estudos foram feitos na área de previsão de *churn*, visto que é um assunto de alto interesse econômico, e tem ganhado tração nas últimas décadas (AHMED et al., 2017). Embora existam diversos levantamentos de dados sobre o assunto (AHMED et al., 2017; BANDARA; PERERA; ALAHAKOON, 2013; AHN et al., 2020), Ahmed et al. (2017)

realizou um estudo completo sobre os conjuntos de dados usados nas previsões de *churn*, dos anos 2000 até 2015, levantando atributos comuns nesses conjuntos de dados, além dos algoritmos mais frequentemente utilizados.

De acordo com Ahmed et al. (2017), os conjuntos de dados utilizados para *churn* englobam, como um todo, atributos como dados pessoais, informações de cobrança e detalhes de contato com os usuários. Tais atributos foram separados pelo autor:

- Informações demográficas;
- Contato ao suporte;
- Informações da conta;
- Dados de compra e cobrança; e
- Relação com o cliente.

Dentro dos dados demográficos, conforme aponta Ahmed et al. (2017), são inclusos dados como gênero, idade, ocupação e local. Os dados de uso do serviço incluem a frequência de uso, duração e categoria das ligações às equipes de suporte das diferentes empresas. Tais atributos eventualmente aparecem em suas formas de média ou soma, caracterizando uma maneira de enriquecer o conjunto de dados (*feature engineering*).

As informações geralmente coletadas sobre a conta do cliente são os tipos de planos e serviços adotados, enquanto as informações de compra e cobrança são relativas a frequência de uso e incidência de pagamento. Dados como valores de conta, *score* de crédito e informações gerais sobre eventuais atrasos também são inclusos nessa categoria (AHMED et al., 2017).

Por fim, dentro da categoria de relação com o cliente, estão dados detalhados sobre a fidelidade, *churns*, mudanças de planos e serviços (AHMED et al., 2017).

Ahmed et al. (2017) também aponta diferentes conjuntos de dados públicos e privados usados para a previsão de *churn*. Dentre os conjuntos de dados públicos destacam-se o *"Churn prediction Dataset"* (JAFARI et al., 2020), disponível no repositório de aprendizado de máquina da Universidade da California (UCI)[1], o conjunto de dados do CRM Teradata, disponibilizado pela Universidade de Duke[2]. Além disso, foi apontado o conjunto de dados *KDD Cup 2009*[3] como um conjunto de dados com muita adesão na literatura (AHMED et al., 2017).

[1]https://archive.ics.uci.edu/ml/
[2]https://www.fuqua.duke.edu/faculty-research/centers
[3]http://kdd.org/kdd-cup/view/kdd-cup-2009/Data

O autor caracterizou as técnicas utilizadas para previsão de *churn* em 5 diferentes categorias, as quais podem ser vistas na Tabela 3.1.

Tabela 3.1: Técnicas e algoritmos frequentemente usados em previsão de *churn*.

Categoria	**Descrição**	**Técnicas**
Previsores tradicionais individuais	Artigos que adotam modelos clássicos individuais comumente usados na pesquisa	Bayesianos, máquinas de vetores de suporte, redes neurais, árvores de decisão
Ensemble de preditores	Artigos que usam da combinação de dois ou mais previsores clássicos	*Bagging*, *Boosting* e modelos híbridos.
Preditores por indução de regra	Artigos que usam indutores para a geração de regras de classificação	RIPPER, C4.5
Preditores com amostragem e *feature engineering*	Artigos com uso de técnicas para lidar com o desbalanceamento das bases de *churn* e gerar novos atributos	SBC, LSE, SIE
Abordagens miscelâneas	Artigos com previsores que fazem uso de atributos não usuais	N/A

Fonte: Ahmed et al. (2017)

Para a validação dos resultados obtidos, o autor aponta que os experimentos costumam fazer uso de *holdout* em diferentes proporções, variando de 50% a 80% da base como treino. O autor aponta também o uso de validação cruzada, apontando 5 como um número de *folds* comumente usado (AHMED et al., 2017).

Em um caso particular nas telecomunicações de uma empresa *churn* na Irlanda (HUANG; KECHADI; BUCKLEY, 2012), notou-se a oportunidade para implantação de previsão de *churn*, com o objetivo de melhorar os serviços e retenção de clientes. Para tal, o autor levantou um conjunto de atributos a partir dos sistemas da empresa de telecomunicação em questão, incluindo dados demográficos, descontos, informações da conta do cliente, ordens de serviço, reclamações e histórico de pagamentos.

Após a geração de atributos e a remoção de *outliers*, os dados foram normalizados e inseridos nos diferentes algoritmos classificadores elencados pelo autor, os quais incluem Regressão Logística (BERKSON, 1944), *Random Forest*(BREIMAN, 2001), Árvores de

decisão (QUINLAN, 2014) e SVM (PLATT, 1998; CORTES; VAPNIK, 1995) (HUANG; KECHADI; BUCKLEY, 2012). Os dados dos 827.124 clientes foram divididos em 50% para treino e 50% teste, os quais continham 738 atributos cada. O autor atingiu valores de ROC-AUC de cerca de 80% nos algoritmos SVM e árvore de decisão (HUANG; KECHADI; BUCKLEY, 2012).

3.2 Interpretação de Resultados de Preditores de *Churn*

Embora o uso prévio de técnicas para interpretação, ou explicação, de modelos de aprendizagem de máquinas aplicados à evasão estudantil sejam desconhecidos durante o momento da escrita deste documento, o uso de tais técnicas em outras áreas ou em problemas similares já existe na literatura.

Em um problema de *churn* de clientes (VILLARREAL et al., 2020) de bancos, existiam uma série de hipóteses iniciais sobre quais fatores que impactavam negativamente e positivamente na experiência do cliente por parte do time de negócio.

Para abordar a situação, Villarreal et al. (2020) utilizou um conjunto fechado de dados bancários com dados comumente encontrados em previsão de *churn*, como idade, local de moradia e emprego. Além disso, o conjunto de dados conta com informações de *feedback* direto do cliente quanto aos serviços fornecidos, como velocidade de atendimento e ambiente agradável, os quais recebiam uma nota de 0 até 10, do menos até mais satisfeito. Nesse contexto, o autor assume a nota 4 como algum grau de insatisfação, as quais são levantadas posteriormente como hipóteses do problema de *churn* (VILLARREAL et al., 2020).

As hipóteses de maior destaque, de acordo com Villarreal et al. (2020), são a distância até a residência, a velocidade de atendimento e o ambiente agradável. Uma análise exploratória foi realizada para melhor elucidar o corportamento de tais variáveis.

Foi constatado que os clientes desertores, quando reclamavam da longa distância até o banco, eram servidores de empresas privadas ou governamentais. Ao analisar a velocidade de atendimento, verificou-se que a maioria dos desertores possuiam idade menor que 40 anos. Por fim, o autor constatou que os clientes que reclamavam de ambiente agradável pertenciam a área de negócios e empresas privadas (VILLARREAL et al., 2020).

Para realizar a previsão de *churn*, o autor realizou testes com diversos algoritmos, incluindo Regressão Logística (BERKSON, 1944), Árvores de decisão (QUINLAN, 2014), *Random Forest* e XGBoost (CHEN; GUESTRIN, 2016), sendo o último o algoritmo com melhor *F-score* (84%) no experimento. Para validação dos resultados, os autores utilizaram da técnica de *hold-out*, com 67% dos dados para treino e 33% para teste.

Após o treino de modelos de previsão de *churn* foi utilizado do SHAP (LUNDBERG; LEE, 2017) para detectar a participação de cada atributo do conjunto de dados e foi possível validar as hipóteses iniciais dos autores. A Figura 3.1 (VILLARREAL et al., 2020) ilustra um exemplo obtido na classificação de um cliente:

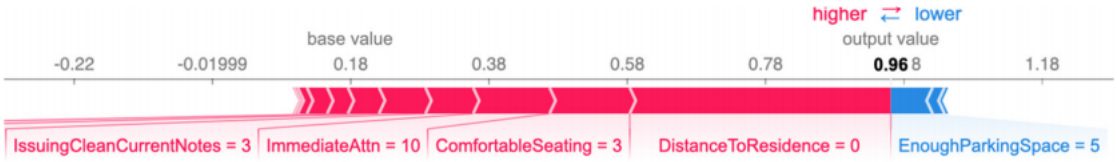

Figura 3.1: Resultados obtidos pelo SHAP para a classificação de um cliente.

Fonte: Villarreal et al. (2020)

O gráfico apresentado na Figura 3.1 indica que a distância até a residência, assentos confortáveis, atendimento imediato tem participação positiva na saída do SHAP, enquanto espaço no estacionamento tem impacto negativo. No caso da classificação de clientes entre *churners* e *non-churners* apresentado por Villarreal et al. (2020), possuir um valor positivo indica que o cliente é menos propenso a mudar de serviço. Além disso, pode-se notar que a distância até a residência é um fator de grande impacto na hora de um cliente permanecer utilizando do banco (VILLARREAL et al., 2020).

Sendo assim, por meio da visualização gráfica gerada pelo SHAP (LUNDBERG; LEE, 2017), foi possível constatar e validar a hipótese criada pelo time de negócio que existia grande participação da distância de moradia com a permanência ou desistência do cliente quanto ao serviço oferecido (VILLARREAL et al., 2020).

Em um caso de uma empresa de telecomunicações romena, DUMITRACHE, NASTU e STANCU (2020) optou por aplicar técnicas de explicação de resultados para melhor compreender os fatores que levam ao *churn* dos clientes da empresa em questão.

Primeiramente, DUMITRACHE, NASTU e STANCU (2020) levantou os dados demográficos, dados de assinatura, descontos e ofertas, poder aquisitivo, informações de pagamento ou atrasos e interações com o suporte da empresa de 10701 clientes.

Após a geração da base de treino, com 75% dos clientes, e da base de teste, com 25%, foram treinados diversas *Random Forests*, as quais obtiveram ROC AUC de 72.6%. Com os preditores devidamente treinados, o autor utilizou o SHAP (LUNDBERG; LEE, 2017) para extrair quais atributos impactavam mais na evasão dos clientes (DUMITRACHE; NASTU; STANCU, 2020). O resultado pode ser visto na Figura 3.2.

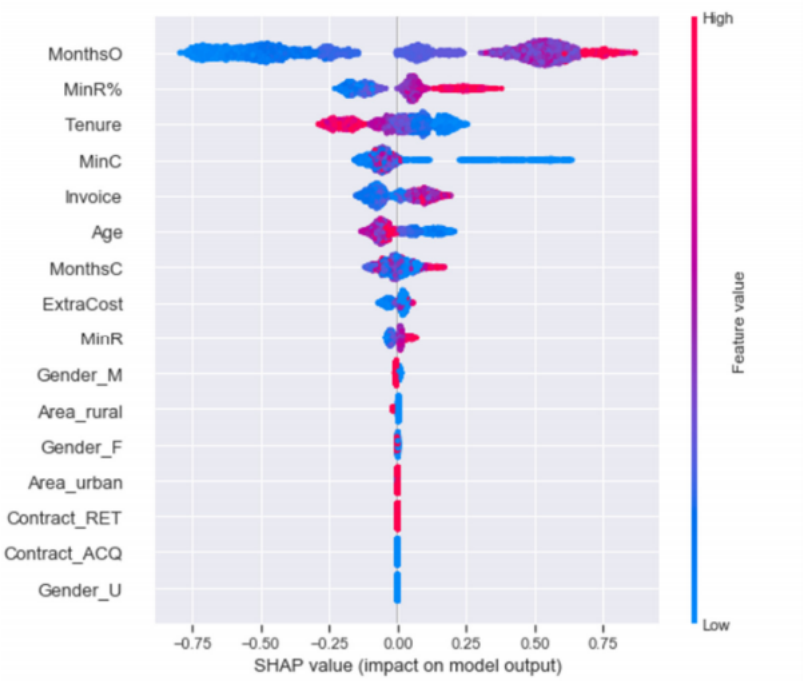

Figura 3.2: Resultados obtidos pelo SHAP para uma base de clientes romena.

Fonte: DUMITRACHE, NASTU e STANCU (2020)

Pode-se constatar que os fatores de maior impacto de acordo com a DUMITRACHE, NASTU e STANCU (2020) são os atributos *MonthsO*, que é a quantidade de meses desde a última oferta, o *MinR%*, que é o porcentagem de minutos em ligação com outras operadoras e a utilização da linha são fatores que levam ao *churn*. Isso pode ser observado pelo gráfico, onde as cores quentes (rosa, vermelho) indicam um alto impacto na deserção, enquanto as cores frias indicam um impacto baixo. Portanto, por meio do SHAP (DUMITRACHE; NASTU; STANCU, 2020), DUMITRACHE, NASTU e STANCU (2020) pode concluir que não realizar ofertas aos clientes, possuir clientes se comunicando muito com clientes de outras operadoras e clientes com pouca utilização são mais prováveis a deserção.

3.3 Fatores de impacto na evasão estudantil

Existem diversos estudos procurando entender os fatores de maior impacto na evasão estudantil, seja do ponto de vista psicológico (MIWA et al., 2015), educacional (BARDACH et al., 2019) ou estatístico (QUARTI; FIGINI; GIUDICI, 2009).

Em um estudo na Universidade da Pavia (QUARTI; FIGINI; GIUDICI, 2009), na Itália, foram analisados os fatores que levam os estudantes de Psicologia e Biologia a desistirem do curso. A escolha do curso de Psicologia se deve à sua dificuldade, quanto

a escolha de Biologia é devida ao fato que muitos estudantes ingressam na esperança de mudar para o curso de Medicina no ano seguinte (QUARTI; FIGINI; GIUDICI, 2009).

O estudo realizado por Quarti, Figini e Giudici (2009) conta com a distinção entre estudantes de evasão voluntária e involuntária, além de estudantes ativos (cursando alguma disciplina) ou inativos (matriculados porém não estão cursando disciplinas).

Foi realizado um levantamento de dados de 845 estudantes de Biologia e 1037 estudantes de Psicologia, dos anos 2001 até 2007. Os atributos levantados incluem identificador estudantil, data de nascimento, gênero, província, tipo de Ensino Médio (regular, técnico, letras, científico), *status* do estudante (ativo ou inativo), total de créditos, inadimplência e média global.

Após uma análise da distribuição das diferentes variáveis coletadas, os autores utilizam de técnicas de Análise de Sobrevivência, fazendo uso da função de Kaplan-Meier (KAPLAN; MEIER, 1958) para estimar a sobrevivência. Os autores concluem que fatores geográficos (localização da moradia original do estudante) e o tipo de ensino médio tem impacto na evasão.

Um estudo desenvolvido na Áustria (BARDACH et al., 2019) em uma instituição do ensino superior, avaliou, usando técnicas de correlação de variáveis, as intenções de evasão de 432 estudantes de mestrado e teve como objetivo avaliar se fatores motivacionais e de contexto na qual o estudante está inserido podem impactar na evasão do alunos do programa. Para tal, foram coletadas informações do perfil dos aluno, como competitividade, independência, busca por melhoria e dedicação. Essas informações foram coletadas por meio de questionários respondidos pelos próprios alunos (BARDACH et al., 2019), o qual atribuiram uma nota de 1 a 6 em diferentes perguntas.

No estudo é constatado que a motivação e metas do estudante, além de contexto que o estudante está inserido, são motivadores para o abandono escolar. Embora o estudo apresente um método robusto, ele é de difícil replicação em instituições com um grande número de alunos visto que exige aplicação de questionários e entrevistas, os quais demandam muito tempo. De tal maneira, é de interesse o desenvolvimento de um método onde seja possível obter tais interpretações em um ambiente de mais fácil reprodução.

3.4 Modelos de previsão de Evasão Estudantil

O uso de técnicas de aprendizado de máquina para prever evasão estudantil já existe na literatura, fazendo uso de diversas técnicas de classificação para abordar esse problema durante o ano letivo (MÁRQUEZ et al., 2016; LYKOURENTZOU et al., 2009; SALES; BALBY; CAJUEIRO, 2016).

Em um estudo realizado no México (MÁRQUEZ et al., 2016), obteve-se dados de estudantes de Ensino Médio, com os quais tinha-se o objetivo de prever evasão estudantil nessas escolas. Foram coletados de 419 estudantes no México. A abordagem proposta pelo autor foi a de realizar a previsão em 7 momentos diferentes do semestre. A tabela com as colunas e momentos de previsão pode ser vista na Tabela 3.2.

Tabela 3.2: Etapas de previsão de evasão propostas.

Momento	**Número de atributos**	**Resumo dos atributos**
1	2	Média no ensino fundamental e média no exame do ensino médio
2	10	Idade, tamanho da sala de aula, presença durante os períodos, renda familiar, bolsa, emprego, educação dos pais
3	11	Deficiência física, altura, medidas corporais, desempenho físico em atividades
4	4	Presença, nível de tédio, advertências e mal comportamento
5	26	Número de amigos, hábitos de estudo, motivação, religião, interesse nas disciplinas, distância da escola, qualidade da infraestrutura escolar
6	7	Notas nas disciplinas
7	1	Evadiu no próximo semestre

Fonte: Adaptado de Márquez et al. (2016)

Os dados usados são de difícil coleta, visto que dependem de provas, exames e questionários aos alunos e pais, mas ao realizar esse esforço durante a duração do semestre, torna-se maior a viabilidade. A Figura 3.3 ilustra os momentos de coleta dos dados (MÁRQUEZ et al., 2016).

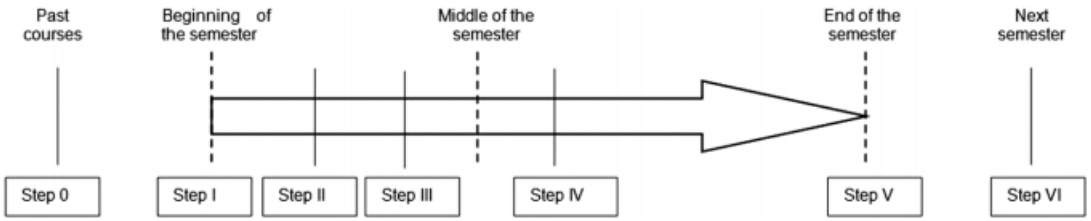

Figura 3.3: Momentos de coleta dos dados.

Fonte: Márquez et al. (2016)

A abordagem proposta por Márquez et al. (2016) consiste em dividir o semestre em vários momentos, espaçados dentro desse período, e coletar os atributos conforme eles forem ficando disponíveis, com a evasão sendo coletada no final do semestre.

Para o protocolo experimental o autor utilizou validação cruzada, adotando como 10 número de divisões. Como algoritmos classificadores, o autor utiliza classificadores como SVM (CORTES; VAPNIK, 1995; PLATT, 1998), Árvore de decisão (QUINLAN, 2014) e PESFAM(LIN et al., 2003).

Após a coleta dos dados, obteve-se um conjunto de dados desbalanceado, com 57 alunos evasores e 362 alunos não evasores. Márquez et al. (2016) fez a análise de 3 experimentos. O primeiro utiliza todos os atributos disponíveis até o momento, no qual o autor obteve valores de *G-mean* variando de 26,4% no começo do curso e alcançando 96.3% no final (MÁRQUEZ et al., 2016).

No segundo experimento o autor utiliza diferentes técnicas para seleção de atributos (MÁRQUEZ et al., 2016), como ganho de informação e valores de chi-quadrado, para selecionar os 10 melhores atributos em cada momento de treino. Nessa abordagem, o autor atinge valores de *G-mean* de até 97,1% por meio do *Naive Bayes*.

Para o terceiro experimento, o autor faz uso de algoritmos para lidar com desbalanceamento de bases. Utilizando o SMOTE (CHAWLA et al., 2002), o autor obteve valores de *G-mean* de até 98,7% (MÁRQUEZ et al., 2016).

Em um estudo realizado em cursos online na área da computação, Lykourentzou et al. (2009) tiveram como objetivo prever a evasão dos alunos combinando diferentes algoritmos classificadores. O estudo conta com um conjunto de dados fechado da Universidade de Atenas, na Grécia. O conjunto contava com dados de 193 alunos, num decorrer de 3 anos e conta com os atributos descritos na Tabela 3.3.

Lykourentzou et al. (2009) separam os atributos em invariantes no tempo, os quais são específicos ao aluno, e variantes, que são relativos às notas nas atividades. Após a coleta dos dados, os autores treinam um *ensemble* de redes neurais, SVM (CORTES; VAPNIK, 1995; PLATT, 1998) e PESFAM (LIN et al., 2003). Os autores realizaram

Tabela 3.3: Atributos adotados para previsão.

Tempo	**Categoria na literatura**	**Atributo**	**Valores Possíveis**
Invariante	Demográfico	Gênero	Homem, Mulher
		Residência	Capital, província
		Experiência de trabalho	Valor em anos
	Experiência	Nível educacional	Básico, Médio, Superior, Mestrado, PHD
		Fluência em inglês	Básico, Intermediário, Alto, Fluente
Variante	Nota na prova de múltipla escolha		Valor entre 0 e 20
	Nota do Projeto		Valor entre 0 e 100
	Dias de atraso		Valor numérico ≥ 0
	Atividade		Valor numérico ≥ 0

Fonte: Adaptado de (LYKOURENTZOU et al., 2009)

a predição em diferentes momentos do curso e obtiveram precisão próxima à 85%. O método utilizado para avaliar os resultados foi o *hold-out*, utilizando 85% dos dados para treino e 15% dos dados para teste.

Dentro da educação brasileira, foi realizado um estudo encima da educação superior da Universidade Federal de Campina Grande (SALES; BALBY; CAJUEIRO, 2016). O conjunto de dados utilizado conta com dados de entrada e saída prevista do aluno, notas na disciplina, dados demográficos do aluno, créditos da disciplina e histórico do aluno com a disciplina. Os dados coletados advém de 32.342 estudantes. Os dados utilizados são referentes aos anos 2002 até 2014.

Após a obtenção da base, o autor define como protocolo de avaliação o uso de treino em semestres anteriores para validar nos posteriores, testando a capacidade de inferência dos modelos. O autor faz uso também de OSS (KUBAT; MATWIN et al., 1997), uma técnica de *undersampling* para contornar o problema de desequilíbrio da base: apenas 61,6% dos estudantes nunca desistiram de uma matéria durante a graduação.

Sales, Balby e Cajueiro (2016), por meio do uso de *Random Forest* (BREIMAN, 2001), obtiveram precisão de cerca de 80%, *recall* de 60% e *F-score* de até 80% nos seus melhores casos dentro da tarefa de previsão de evasão. O autor também destaca a importância do uso do *F-score* para a classificação em problemas com desbalanceamento de classes (SALES; BALBY; CAJUEIRO, 2016). O estudo não explora a questão dos motivos que levam a evasão, mas ilustra as técnicas que podem ser usadas para a previsão da mesma.

3.5 Considerações Finais

A previsão de evasão é um tópico já abordado na literatura, embora a interpretação dos modelos de previsão não seja ainda abordada. Entretanto, a interpretação de modelos já é explorada em problemas com definição similar. Uma delas é a área de previsão de *churn*, na qual a telecomunicação se destaca como expoente (AHMED et al., 2017). É possível verificar que existem diversos fatores, como dados demográficos e situação financeira por exemplo, que afetam a permanência de um cliente em um serviço, e tais fatores são também aplicáveis na previsão de evasão (QUARTI; FIGINI; GIUDICI, 2009).

Dados demográficos, de satisfação, informações de cobrança e inadimplência são fatores comuns que afetam a permanência de um cliente em um serviço, os quais são comumente usados em algoritmos de aprendizado de máquina para a previsão de *churn*. As metodologias de treino e teste variam entre *hold-out* e validação cruzada, mas é constante o uso de classificadores monolíticos, como SVM (PLATT, 1998; CORTES; VAPNIK, 1995) e árvores de decisão (QUINLAN, 2014), e *ensembles*, como XGBoost (CHEN; GUESTRIN, 2016) e *Random Forest* (BREIMAN, 2001). Dentro desses, os *ensembles* apresentaram melhores resultados quando avaliados em conjunto com os demais algoritmos.

Devido ao maior número de artigos publicados e maior interesse econômico por parte das empresas, já se faz uso de técnicas como o SHAP (LUNDBERG; LEE, 2017) para explicação dos resultados obtidos pelos classificadores. O SHAP possibilitou validar os parâmetros de maior contribuição nas diferentes situações onde foi aplicado, o que possibilitou a atuação das equipes envolvidas para melhorar a qualidade geral dos serviços (VILLARREAL et al., 2020).

Na tarefa de previsão de evasão, o retrato é similar ao *churn* quanto a coleta de dados. Dados demográficos são comumente utilizados também, bem como os de aderência aos serviços da escola. Entretanto, dados como notas e faltas são exclusivos à tarefa de previsão de evasão, os quais são amplamente usados na literatura.

As técnicas de avaliação utilizadas para verificar os resultados também são similares ao *churn*, onde *hold-out* e validação cruzada são costumeiramente usados. Além disso, também se faz uso de algoritmos monolíticos e *ensembles* para a previsão. É comum também utilizar técnicas para lidar com o desequilíbrio que geralmente ocorre nas bases de evasão.

Levando em consideração os pontos anteriormente levantados, pode-se constatar que existem casos de sucesso na literatura para a previsão de evasão (QUARTI; FIGINI; GIUDICI, 2009; BARDACH et al., 2019; MÁRQUEZ et al., 2016). Entretanto, grande parte dos conjuntos de dados são de difícil reprodutibilidade, o que leva aos *datasets*

abrangirem um pequeno número de alunos. Além disso, o tema de interpretação de modelos no contexto de previsão de evasão não é comumente abordado na literatura. Tema que apresentou contribuições em áreas correlatas, como o *churn*. De tal maneira, a criação de uma abordagem de maior reprodutibilidade para previsão de evasões, contando com interpretações para melhor entender o processo de decisão dos modelos é justificada.

Capítulo 4

Metodologia de Pesquisa

Este trabalho conta como objetivo a apresentação de uma metodologia para desenvolvimento de modelos de aprendizagem de máquina para a tarefa de predição de evasão de alunos. Para tal, pretende-se construir e avaliar indutores que, além da detecção da potencial evasão, forneçam uma explicação para o motivo desta por meio de técnicas de interpretação de modelos. A fim de atingir os objetivos mencionados, pode-se dividir o problema em diversas etapas e abordá-las individualmente, as quais a metodologia proposta na Figura 4.1 demonstra.

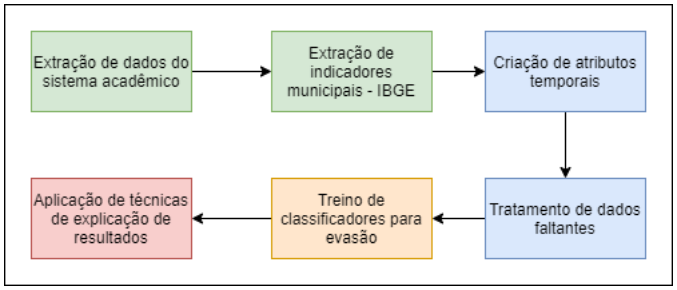

Figura 4.1: Proposta de metodologia.

Fonte: O autor

Inicialmente, faz-se necessário o entendimento, levantamento e extração dos dados dos alunos (como por exemplo notas, valores de anuidade e localidade) a partir do sistema acadêmico. A Seção 4.1 deste capítulo trata do processo desenvolvido para a coleta dos dados e geração do conjunto de dados inicial.

Após o levantamento dos dados básicos dos alunos, torna-se possível enriquecer a base com dados externos relevantes, como por exemplo, o PIB e expectativa de anos de estudo fazendo uso dos dados fornecidos pelo IBGE. A Seção 4.2 deste capítulo trata do processo para obtenção dos dados de indicadores regionais adotados na confecção do conjunto de dados.

Como o ano letivo é dividido em diversos trimestres, é possível fazer a coleta de informações dos alunos em cada trimestre. Para incorporar essas características temporais, existe a possibilidade do uso de técnicas de *feature engineering* para enriquecer a base com atributos que melhor descrevam as variações durante o tempo e o status geral. A Seção 4.3 deste capítulo detalha as definições de fórmulas e atributos criados para enriquecer a base com características temporais.

Tendo uma base de dados levantada e devidamente enriquecida, é possível tratar eventuais falhas na base. Falhas comuns em extrações são, por exemplo, dados faltantes ou fora do valor esperado (*outliers*). A Seção 4.4 aborda as técnicas usadas para tratar os dados faltantes encontrados.

Considerando que a base contém os atributos necessários, e não contém falhas grandes, torna-se possível utilizar algoritmos classificadores para classificar, e assim prever, a evasão dos alunos em diferentes momentos do ano letivo. Além disso, é viável também criar modelos preditivos não apenas para os momentos diferentes do ano letivo mas como para os diferentes segmentos de ensino, levando em consideração que o contexto no qual estão inseridos é diferente. A Seção 4.5 aborda os algoritmos classificadores, técnicas de avaliação e divisões da base adotadas neste trabalho.

Dado o treino dos diferentes modelos preditores de evasão, é possível realizar a interpretação dos mesmos de maneira a gerar potenciais explicações sobre os motivos que levaram um estudante a abandonar a escola. Além disso, é possível obter quais variáveis que apresentaram maior impacto na evasão estudantil para cada segmento de ensino. A Seção 4.6 define como os resultados dos classificadores podem ser utilizados para definir estudantes de risco, enquanto a Seção 4.7 trata de técnicas utilizadas para inferir explicações a partir dos modelos obtidos para previsão de evasão.

4.1 Levantamento da Base de Dados de Alunos

A obtenção de dados de alunos partiu dos dados do Grupo Marista, que possui escolas distribuídas em diferentes lugares no Brasil. Como fonte inicial de dados, foi utilizado o sistema transacional Prime (MANNESOFT, 2021), ferramenta que o Grupo Marista utiliza para a gestão de seus alunos. A base, devido a questões de sensibilidade e segurança, é privada, anonimizada, não possui dados pessoais que permitam identificação direta dos estudantes e a execução de todos os experimentos foi conduzida dentro da rede de computadores institucional.

As informações coletadas consistem das informações de 164,038 alunos, de 22 escolas privadas do Grupo Marista, nos anos de 2015 até 2019 dos segmentos de Educação

Infantil, Educação Básica e Ensino Médio. As informações relativas aos anos letivos são coletadas a partir dos três trimestres do ano letivo.

O sistema Prime conta com diversos dados dos alunos, visto que é usado para fazer o acompanhamento de toda a vida acadêmica nas escolas. Dentro de todas as bases, foi desenvolvida uma consulta, a partir de uma réplica do banco do sistema Prime, para extrair as informações dispostas na Tabela 4.1:

Tabela 4.1: Atributos extraídos do sistema acadêmico.

Atributos extraídos do sistema acadêmico para o conjunto de dados	
Ano escolar	Quantidade de faltas
Escola	Presença de CPF, RG, CTPS cadastrados no sistema
Valor de anuidade	Menor e maior nota no ano
Sexo	Média no ano
Tempo na escola	Participação de atividades extracurriculares
Idade	Profissão e escolaridade dos protetores
Anos de atraso com relação ao esperado para a série	Evasão durante o ano letivo

Além dos dados básicos do alunos, foram levantadas as diferentes notas de cada um dos três trimestres do ano quando disponível. As disciplinas, nas diferentes regiões do Brasil possuem diferentes nomes. Após corretamente mapear as disciplinas para um nome comum entre todos os estados, foi possível escolher quais disciplinas seriam utilizadas neste estudo. As disciplinas elencadas, caso disponíveis, para os diferentes segmentos de ensino estão dispostas na Tabela 4.2.

Tabela 4.2: Disciplinas disponíveis do sistema acadêmico PRIME.

Disciplinas utilizadas			
Artes	Geografia	Português	Filosofia
Biologia	História	Química	Sociologia
Ciência	Inglês	Educação Física	Matemática
Física			

Para a definição da classe alvo foi adotado o critério de quebra de contrato, isto é, um aluno é dado como evasor durante aquele ano letivo caso tenha quebrado o contrato

e saído durante o ano letivo. Casos de transferência interna dentro da rede também são contados como casos de evasão. Foram estudadas outras possibilidades para a definição da classe alvo, como a não-rematrícula no ano seguinte (critério também chamado de evasão dentro das escolas do Grupo Marista) mas foram descartadas por não se adequarem a definição comumente adotada pela literatura (LEON; MENEZES-FILHO, 2002; BARDACH et al., 2019; MIWA et al., 2015). Esses critérios serão adotados como futuros estudos dentro da rede de escolas e não fazem parte do escopo desse trabalho.

4.2 Enriquecimento da Base

Apesar de os dados representarem a situação do aluno na escola, o ambiente externo pode afetar o dia a dia do aluno, e por sua vez influenciar na sua evasão. De tal maneira, é possível usarmos dados com base na moradia do aluno para melhorar o resultado da classificação.

A situação econômica de um município impacta empregos e fluxo de capital. Esses fatores podem impactar, indiretamente, na vida dos alunos e suas famílias. Portanto, é intuitivamente vantajoso adicionar tais informações ao conjunto de dados.

Os dados de moradia do aluno foram então cruzados com as informações na API dos correios (CORREIOS, 2021), a qual permite obter, o nome do município e o estado onde se localiza. Isso se faz necessário devido ao fato de existirem alunos que não moram no município onde estudam. Os dados de município foram então cruzados com as bases de indicadores do IBGE (IBGE, 2021), da qual foram extraídas as características descritas na Tabela 4.3.

Tabela 4.3: Dados municipais extraídos da base do IBGE.

Dados municipais extraídos da base do IBGE	
Código do IBGE do município	Índice de frequência escolar municipal
PIB	Coeficiente GINI
IDHM	Proporção de pobreza extrema
IDHME	Proporção de pobreza extrema infantil
IDHML	Expectativa de vida
IDHMR	Taxa de fecundidade
Índice de escolaridade	Expectativa de anos de estudo

Com os dados devidamente coletados, foi utilizado Apache Spark (ZAHARIA et al., 2016) para realizar a junção das informações de alunos e municípios.

4.3 Criação de Atributos Temporais

Os dados iniciais, embora descrevam o aluno, não abrangem suficientemente a variação da performance durante o ano, fator que pode afetar a permanência na escola. Sendo assim, é importante fornecer essa informação para os diferentes algoritmos classificadores para que possam possuir uma melhor performance.

Levando esse problema em consideração, foram criados diferentes atributos para melhor descrever tais comportamentos.

4.3.1 Derivada trimestral das notas

Para demonstrar a queda ou melhoria brusca de performance do aluno, foram adicionados atributos de variação de nota entre trimestre. Para cada uma das notas, foi calculada a variação da mesma nota para o trimestre anterior conforme a fórmula abaixo (onde N_i é a nota N no i-ésimo trimestre):

$$\Delta N = N_t - N_{t-1} \tag{4.1}$$

Em casos onde não existe histórico trimestral, a nota anterior assume valor 0. Este preenchimento é dado devido a alguns segmentos não possuírem todas as disciplinas contempladas no conjunto de dados, ou ainda alguma escola não ofertar a disciplina em questão.

4.3.2 Derivada anual das notas

Da mesma maneira, é possível calcular a variação entre os diferentes anos (em caso de alunos que estejam a vários anos na escola) e demonstrar se houve uma queda ou melhoria entre os diferentes anos escolares. Para cada uma das notas, foi calculada a variação da mesma nota para o ano anterior conforme a fórmula abaixo (onde N_i é a nota N no i-ésimo ano):

$$\Delta N = N_a - N_{a-1} \tag{4.2}$$

Em casos onde não existe histórico anual, a nota anterior assume valor 0. Este preenchimento é dado devido a alguns segmentos não possuírem todas as disciplinas contempladas no conjunto de dados, ou ainda alguma escola não ofertar a disciplina em questão.

4.3.3 Nota acumulada durante o ano

Para demonstrar a performance anual, e mitigar efeitos de eventuais quedas ou aumentos repentinos, é possível criar uma *feature* que corresponda ao progresso da nota durante o ano. De tal maneira, a *feature* descrita pode ser descrita pela seguinte fórmula (onde N_i é a nota N no i-ésimo ano e T é o total de trimestres):

$$N_{acumulada} = \sum_{1}^{T} N_i \tag{4.3}$$

Em casos onde não existe histórico trimestral, a nota anterior assume valor 0. Este preenchimento é dado devido a alguns segmentos não possuírem todas as disciplinas contempladas no conjunto de dados, ou ainda alguma escola não ofertar a disciplina em questão.

4.3.4 Exemplo prático dos atributos derivados

Para melhor compreensão da abordagem aplicada, a Tabela 4.4 ilustra um exemplo de como esses novos atributos derivados apresentam-se no conjunto de dados.

Tabela 4.4: Exemplo contendo os atributos derivados.

Identificador do estudante	**Ano letivo**	**Trimestre**	**N**	$\mathbf{N_{acumulado}}$	**ΔN**
123456789	2015	1	7.9	7.9	0
123456789	2015	2	6.5	14.4	-1.4
123456789	2015	3	3	17.4	-3.5
123456789	2016	1	8.1	8.1	0

Representado na Tabela 4.4, um estudante fictício, identificado por 123456789, tem suas notas N para diferentes trimestres em 2015 e 2016. Pela equação 4.3, o valor na coluna $N_{acumulado}$ representa a soma de todas as notas até o trimestre da linha no respectivo ano escolar.

Da mesma maneira, a coluna ΔN, dada pela Equação 4.1 representa a diferença entre as notas para o estudante nos diferentes trimestres do ano letivo.

É importante notar que a computação desses valores foi feita para cada ano letivo individualmente, não levando em consideração os anteriores. visto que novos estudantes em uma escola não teriam valores para tais variáveis.

4.4 Tratamento de Características Faltantes

Dada a origem dos valores ser um sistema transacional, e este estar suscetível à erros e à mudanças, é possível que nem todos os dados estejam preenchidos ou corretos. Isso ocasiona, ao gerar o conjunto de dados, valores faltantes.

A existência de dados faltantes não é prevista em muitos algoritmos de classificação. Para resolver este requisito de não-absenteísmo de valores, podem ser empregadas diferentes técnicas para preencher os valores. Para as colunas onde os dados são numéricos, isto é, podem assumir um valor dentro de um intervalo numérico. O uso da mediana é um dos métodos mais simples e apresenta bons resultados na maior parte dos conjuntos de dados (JADHAV; PRAMOD; RAMANATHAN, 2019). Para as colunas onde os dados são categóricos, ou seja, assumem um valor de um número fixo de possíveis valores, foi inserido um arbitrário pré-determinado, i.e., `NA`, servindo como representação de valor faltante, de modo a não mascarar a falta desses valores.

Foi realizada, após a geração da base inicial, a agregação de novos atributos e o tratamento das características faltantes uma análise sobre o conjunto de dados, elencando número de valores não-nulos e tipos dos dados além da tradução dos nomes dos atributos em código para uma linguagem facilmente compreensível por leitores externos.

4.4.1 Divisão da base e medidas de precaução contra *leaks*

Para a criação de uma *baseline*, a primeira rodada de modelos de previsão de evasão foi feita sob a base completa, isto é, sem segmentação por trimestre ou segmento estudantil. Isto é feito de maneira a gerar uma *baseline* para comparação com os demais modelos a serem treinados.

Entretanto, devido à alta complexidade do problema, seja pela sazonalidade dos dados ou por diferentes perfis comportamentais no conjunto de dados, é possível abordar segmentos do *dataset* com diferentes técnicas com a finalidade de obter modelos melhor ajustados às necessidades do segmento em questão. De tal maneira, foram realizadas divisões no conjunto de dados, abordando combinações de segmentos estudantis e trimestres separadamente para a posterior classificação.

Para evitar potenciais *leaks* de dados é necessária a tomada de medidas preventivas. Caso um aluno saia antes do fechamento de um trimestre, o mesmo não aparecerá nos subsequentes trimestres, exceto em casos onde haja notas parciais referente ao próximo trimestre. Além dessa medida, o identificador do aluno não é dado como entrada ao modelo, evitando assim a identificação do aluno pelo classificador, mesmo que o mesmo

apareça em instâncias de anos seguintes. Por fim, o protocolo de validação cruzada, o qual será comentado adianta, não permite que uma mesma instância seja usada como teste e treino na mesma rodada de treinos, característica que também previne *leaks*.

4.4.1.1 Segmentos de ensino

O Grupo Marista possui escolas que atendem desde a educação infantil até o ensino médio. Portanto, o sistema transacional das escolas, o qual foi a origem dos dados, conta com diversos segmentos de ensino. Os segmentos estudantis apresentam diferentes comportamentos e diferentes disciplinas. Por esse motivo, é possível separar os diferentes segmentos de ensino e treinar modelos que levem em consideração suas peculiaridades individualmente. Os diferentes segmentos a serem classificados são:

- Ensino infantil;
- Ensino fundamental - Anos iniciais;
- Ensino fundamental - Anos finais; e
- Ensino médio.

4.4.1.2 Trimestres

Devido à sazonalidade dos dados a base foi dividida, também, entre trimestres. Sendo assim, cada trimestre termina por representar, cada um, cerca de 33% do ano letivo. Essa separação é importante devido ao motivo de evasão poder variar conforme o passar do ano.

Sendo assim, as segmentações disponíveis para treino em diferentes momentos do ano letivo são as seguintes:

- Primeiro trimestre;
- Segundo trimestre; e
- Terceiro trimestre.

Levando em consideração que questões como greves, feriados regionais e leis estaduais podem impactar o ano letivo, a definição das datas que definem os trimestres é definida individualmente por cada escola. Sendo assim, não há padrão definido ou critério para a definição de uma data de limite ou data de corte.

É importante notar que embora as combinações de segmentos estudantis e trimestres tenham sido analisados separadamente, foi notado que o comportamento perante aos trimestres foi similar entre todos eles. Isto é, a performance dos modelos mudou de maneira similar com o passar dos pontos de cortes temporais independente do segmento estudantil escolhido. Sendo assim, para a demonstração dos resultados obtidos pela divisão trimestral neste trabalho, foi realizado o agrupamento independente do segmento estudantil de todos os alunos disponíveis para cada trimestre. Essa decisão foi tomada de maneira a melhor mostrar o comportamento com o passar do ano letivo, em contrapartida a mostrar variações menores específicas de cada segmento de ensino.

4.5 Treino de classificadores para previsão de evasão

Com a base devidamente tratada e as divisões estabelecidas, é possível trabalhar na questão de previsão de evasão. De tal maneira, é necessário testar diferentes algoritmos classificadores e diferentes parâmetros de inicialização.

Para a definição dos hiper-parâmetros dos algoritmos, foi utilizado o Azure AutoML (FUSI; SHETH; ELIBOL, 2018) como ponto de início para os parâmetros, os quais foram posteriormente otimizados manualmente.

Para a determinação do melhor modelo preditor de evasão, foram testados diversos algoritmos classificadores caixa branca e caixa preta. Isso foi feito de maneira a selecionar os que apresentaram os melhores resultados em cada categoria, tendo assim um equilíbrio entre modelos nativamente interpretáveis e modelos de mais complexa interpretação.

De tal maneira, os classificadores selecionados para o processo final de classificação foram:

- AdaBoost (FREUND; SCHAPIRE, 1997)
- Regressão Logística (BERKSON, 1944)
- XGBoost (CHEN; GUESTRIN, 2016)
- Random Forest (BREIMAN, 2001)
- Árvore de Decisão (QUINLAN, 2014)

As implementações adotadas são as presentes na biblioteca Scikit-Learn (PEDREGOSA et al., 2011), exceto no caso do XGBoost, o qual conta com biblioteca própria (CHEN; GUESTRIN, 2016).

É importante notar que embora abordados na literatura, algoritmos caixa preta como Máquinas de Vetor de Suporte e Redes Neurais não serão abordados nos resultados obtidos. Isso é devido a piores resultados apresentados nos testes iniciais quando comparados aos demais algoritmos caixa preta avaliados.

Para a obtenção dos resultados foi adotada a validação cruzada, utilizando como número de *folds* 5, onde a classe (evasor ou não-evasor) foi usada para a estratificação. A metodologia de divisão da base para criação dos subconjuntos de treino e teste, os quais são usados para alimentar a validação cruzada, pode ser visualizada na Figura 4.2.

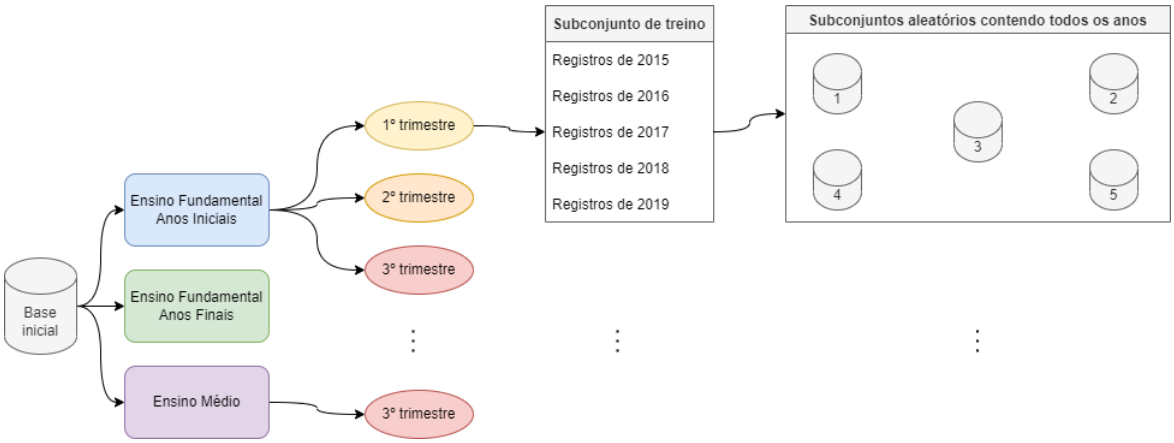

Figura 4.2: Metodologia para divisão da base em subconjuntos.

Fonte: O autor

A adoção da metodologia de validação cruzada não permite que uma mesma instância seja usada tanto como treino e teste em uma rodada de treino, evitando *leaks*. As métricas avaliadas para a comparação dos classificadores nesse protocolo foram:

- Área sob a curva precisão-revocação (PR AUC);
- F1-score;
- Área sob a curva característica de operação do receptor (ROC AUC); e
- Acurácia.

A escolha das métricas se deve ao fato de que as três primeiras apresentam diferentes visões sobre conjuntos de dados desbalanceados. Embora não tão robusta à desequilíbrios na classe alvo, a inclusão da acurácia foi dada de maneira a representar o quanto da base geral se teve como acerto.

4.6 Sinalização de estudantes de risco

Com base nas probabilidades de evasão dadas pelos modelos, serão sinalizados como evasores os estudantes que apresentam maiores riscos de saírem da escola. Para sinalização dos estudantes será adotado um critério de alertas, o qual indicará a saúde da estadia do estudante na escola. Isto será feito de modo a simplificar a compreensão por times educacionais, sem o conhecimento técnico e estatístico por trás das previsões.

Para adoção do critério de alerta vermelho, foi levado em consideração o alto tempo necessário para converter uma evasão em permanência, especialmente em uma situação de alta probabilidade desta evasão. Portanto a marcação deve ser feita apenas em casos onde se há maior probabilidade de que o aluno sairá da escola.

Por sua vez, o critério de definição de alerta amarelo leva em consideração o desejo de detectar um grande número de alunos que podem evadir, mas não necessariamente sua evasão é certa no momento da previsão. Por consequência, o tempo necessário para abordar vários alunos de maneira mais sutil é menor, ao se comparar em um caso onde a evasão é altamente provável.

Para a definição desses *thresholds* foi utilizada a curva Precisão-Recall (PR), usando como base os modelos de anos anteriores a cada caso avaliado. Em alguns casos deseja-se maximizar a precisão da informação (um caso de alto risco de evasão) e em outros deseja-se maximizar a revocação, captando o maior número possível de potenciais evasores (vários casos de risco médio de evasão). Portanto a definição do nível de saúde do aluno no alerta será dado pelos seguintes critérios:

- *Alerta Vermelho*: A probabilidade de evasão do aluno cai em um ponto de pelo menos 80% de precisão, tomando com base a curva PR dos modelos dos anos anteriores;

- *Alerta Amarelo*: A probabilidade de evasão do aluno cai em um ponto de pelo menos 60% de precisão e onde a revocação é máxima, tomando com base a curva PR dos modelos dos anos anteriores; e

- *Alerta Verde*: A probabilidade de evasão do aluno não se encaixa em nenhum dos outros casos.

Aliado aos atributos de maior contribuição para evasão na interpretação fornecida pelo SHAP, o sinal de saúde de estadia do estudante potencialmente servirá às equipes educacionais converterem a evasão em permanência, acompanhamento o qual não cabe no escopo deste trabalho.

4.7 Interpretação de resultados de modelos

Dada a etapa de treino dos previsores de evasão e a definição dos riscos de evasão, torna-se possível observar os resultados obtidos pelos classificadores de maneira a extrair explicações dos modelos. Levando em consideração o elevado número de modelos treinados, serão obtidas interpretações apenas a partir dos melhores modelos de cada segmento estudantil.

Para a geração de explicações dos modelos, primeiramente será feita a análise dos atributos que mais tiveram impacto no modelo, a qual é fornecida pelas implementações dos algoritmos adotadas pra o desenvolvimento deste trabalho (PEDREGOSA et al., 2011; CHEN; GUESTRIN, 2016). Serão ilustrados, de maneira gráfica, os atributos que mais contribuem para a performance do modelo, de maneira a evidenciar os fatores de maior impacto em uma permanência ou evasão.

Após a análise dos atributos de maior importância, serão geradas explicações a partir dos modelos treinados. Isso será feito de avaliar quais atributos tem maior participação e importância na classificação de casos gerais e específicos.

Em casos onde os classificadores caixa branca, como a Árvore de Decisão (QUINLAN, 2014), apresentarem melhores resultados, o protocolo de interpretação dos modelos toma como base a representação gráfica nativa do algoritmo. A Figura 2.6, anteriormente apresentada, demonstra uma representação obtida por Pedregosa et al. (2011) ao treinar uma Árvore de Decisão, um algoritmo caixa branca.

Além do uso da representação gráfica dos algoritmos caixa branca, será utilizado o SHAP (LUNDBERG; LEE, 2017) nos algoritmos de caixa-preta, como o XGBoost (CHEN; GUESTRIN, 2016). O uso do SHAP permite a obtenção das participações dos diferentes atributos em casos de classificação mesmo em algoritmos caixa preta.

A Figura 2.11, apresentada anteriormente no Capítulo 2, demonstra o uso do SHAP em um algoritmo caixa-preta.

Conforme demonstra a Figura 2.11, pode-se por exemplo notar que a participação do atributo "*petal length*" é maior do que a dos demais, devido a área que ocupa no gráfico. Além disso, nota-se que para valores menores do que 10, ela contribui para uma classe, em outros casos para outra (LUNDBERG; LEE, 2017). Sendo assim, é possível a extração de interpretações do modelo a partir da representação gráfica do SHAP.

Por meio do uso das visualizações geradas pelo SHAP ou pelas representações dos algoritmos caixa-branca, pretende-se levantar as variáveis com alto impacto na evasão estudantil e gerar hipóteses sobre as razões de tais variáveis ocasionarem ou que reflitam a evasão. Além disso os valores de SHAP serão usados para definir também a nível de aluno, quais fatores estão contribuindo para casos individuais de evasão, ilustrando os

problemas de cada estudante.

Capítulo 5

Resultados

Os resultados deste trabalho estão divididos em três partes. Primeiramente, foi criado um conjunto de dados usando os dados das 22 escolas da região Centro-Sul do Grupo Marista. Este consiste na Seção 5.1 dos resultados. A partir da base, foram treinados múltiplos classificadores para previsão de evasão neste mesmo conjunto de dados. Os resultados da classificação podem ser visualizados na Seção 5.2. Tendo como objeto de estudo o melhor modelo encontrado para cada segmento estudantil, foi realizada na Seção 5.3 a análise dos fatores que levam a evasão escolar, em conjunto das interpretações extraídas a partir dos modelos treinados.

5.1 Conjunto de Dados

Como primeiro resultado deste projeto foi obtida uma base de dados contemplando os dados das 22 escolas da região Centro-Sul pertencentes ao Grupo Marista. Os dados, gerados a partir do ano de 2015 até 2019, englobam alunos desde a Educação Infantil até o Ensino Médio.

5.1.1 Informações gerais do conjunto de dados

O conjunto de dados gerado possui 406,293 linhas, as quais englobam 8,070 casos de evasão nas escolas e 155,968 correspondem a casos onde o ano foi terminado regularmente. De tal maneira, o conjunto de dados pode ser dito como desbalanceado, visto que a proporção entre casos positivos e negativos é alta.

Cada instância no conjunto de dados corresponde a um aluno em um trimestre. O conjunto de dados compreende todos os segmentos da educação básica. A Tabela 5.1 demonstra a distribuição da base dentro de cada segmento educacional.

Tabela 5.1: Distribuição das classes por segmento de ensino.

Segmento	**Evasão**	**% evasão**	**Permanência**	**% permanência**
Infantil	2625	9,66%	24545	90,34%
Fundamental - Anos Iniciais	1840	3,14%	56731	96,86%
Fundamental - Anos Finais	1606	3,38%	45908	96,62%
Médio	1999	6,49%	28784	93,51%
Total	**8070**	**4,92%**	**155968**	**95,08%**

Pode-se notar que o desbalanceamento de classes é alto em todos os segmentos, com o segmento infantil se mostrando aquele com a maior taxa de evasão.

Para fins de consulta, os Apêndices B e C demonstram o tipo de distribuição, valores médios e desvios padrões para as colunas do conjunto de dados gerado dentro de cada segmento e trimestre, respectivamente.

5.1.2 Dados do aluno

O conjunto de dados contém diversas informações referentes à dados de alunos. A Tabela 5.2 contém as colunas e informações coletadas referentes aos dados pessoais dos alunos e tem como origem o sistema transacional do Grupo Marista.

Tabela 5.2: Atributos de dados da matrícula e ano letivo no conjunto de dados.

Nº	**Nome do atributo**	**Não-nulos**	**Tipo**
1	Ano Letivo	100,00%	*integer*
2	Código do Aluno[1]	100,00%	*string*
3	Anuidade	100,00%	*float*
4	Escola[1]	100,00%	*string*
5	Turma[1]	100,00%	*string*
6	Série[1]	100,00%	*string*
7	Segmento Estudantil	100,00%	*integer*
8	Ano de entrada	100,00%	*float*
9	Tempo de escola	100,00%	*float*
10	Idade	100,00%	*float*
11	Anos de atraso	100,00%	*float*
12	Total de descontos	56,78%	*float*
13	Total de faltas	95,45%	*float*
14	Média global histórica	96,03%	*float*

15	Menor nota global histórica	96,03%	*float*
16	Maior nota global histórica	96,03%	*float*
17	Faz atividades complementares	100,00%	*integer*
18	Frequenta ensino integral	100,00%	*integer*
19	Faz atividades religiosas	100,00%	*integer*
64	Código da Escola	100,00%	*integer*
65	Sexo do aluno	100,00%	*integer*
66	Provedor de email da protetora	100,00%	*integer*
67	Provedor de email do protetor	100,00%	*integer*
68	Código do segmento de ensino	100,00%	*integer*
69	Código da série	100,00%	*integer*
114	Aluno Evadiu (target)	100,00%	*integer*

Os dados disponíveis compreendem as informações básicas do aluno e ano letivo. Além disso, agregam informações gerais sobre sua carreira estudantil como quantos anos está atrasado e se frequenta atividades extracurriculares.

5.1.3 Dados da região

O conjunto de dados foi enriquecido com indicadores de desenvolvimento e qualidade de vida da região na qual o aluno mora. A Tabela 5.3 demonstra os indicadores levantados para o uso na tarefa de classificação.

Tabela 5.3: Atributos regionais no conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
49	IDHM da localidade	98,90%	*float*
50	IDHME da localidade	98,90%	*float*
51	IDHML da localidade	98,90%	*float*
52	IDHMR da localidade	98,90%	*float*
53	Índice de escolaridade da localidade	98,90%	*float*
54	Índice de frequência escolar da localidade	98,90%	*float*
55	GINI da localidade	98,90%	*float*
56	Proporção de pobreza extrema da localidade	98,90%	*float*
57	Proporção de pobreza extrema infantil da localidade	98,90%	*float*
58	Proporção de pobreza da localidade	98,90%	*float*

[1]Essa coluna não é utilizada pelos classificadores e é apenas para auxiliar na manipulação da base

59	Expectativa de vida da localidade	98,90%	*float*
60	Fecundidade da localidade	98,90%	*float*
61	Expectativa de anos de estudo da localidade	98,90%	*float*
62	PIB da localidade	98,90%	*float*
63	PIB per capita da localidade	98,90%	*float*

Os dados vem de informações de censos do IBGE do ano mais recente para a região na qual o aluno mora.

5.1.4 Dados dos protetores

A partir do sistema transacional é possível obter as informações básicas dos protetores responsáveis pelos alunos. Embora existam exceções, os protetores cadastrados na grande maioria dos casos são o pai e mãe dos alunos. As informações básicas coletadas estão disponíveis na Tabela 5.4 são referentes à documentos, escolaridade, profissão e provedor de e-mails de ambos os responsáveis pelo aluno.

Tabela 5.4: Atributos dos responsáveis no conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
20	Possui protetor	100,00%	*integer*
21	Possui protetora	100,00%	*integer*
22	Protetor/a faleceu	98,12%	*integer*
23	Religião do protetor	63,86%	*integer*
24	Escolaridade do protetor	60,22%	*integer*
25	Profissão do protetor	81,22%	*integer*
26	Parentesco do protetor cadastrado	49,28%	*integer*
27	Protetor possui CTPS	98,12%	*integer*
28	Protetor possui RG	98,12%	*integer*
29	Religião da protetora	66,52%	*integer*
30	Escolaridade da protetora	62,69%	*integer*
31	Profissão da protetora	81,48%	*integer*
32	Parentesco da protetora cadastrada	50,16%	*integer*
33	Protetora possui CTPS	99,67%	*integer*
34	Protetora possui RG	99,67%	*integer*

Os dados referentes aos protetores foram coletados a partir do sistema transacional do Grupo Marista.

5.1.5 Dados das notas

Para introduzir a performance do aluno no trimestre sendo avaliado, foram adicionadas como *features* as notas nas disciplinas comumente ofertadas nas escolas referentes ao trimestre correspondente à linha em questão. As informações na Tabela 5.5 são referentes às notas trimestrais presentes no conjunto de dados.

Tabela 5.5: Atributos de desempenho do aluno no conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
35	Trimestre	89,77%	*float*
36	N_i Artes	84,17%	*float*
37	N_i Biologia	21,97%	*float*
38	N_i Ciências	67,39%	*float*
39	N_i Educação Física	88,72%	*float*
40	N_i Filosofia	29,81%	*float*
41	N_i Física	21,97%	*float*
42	N_i Geografia	89,07%	*float*
43	N_i História	89,08%	*float*
44	N_i Inglês	86,99%	*float*
45	N_i Matemática	89,11%	*float*
46	N_i Português	89,16%	*float*
47	N_i Química	21,97%	*float*
48	N_i Sociologia	21,59%	*float*

As informações de notas, advindas do sistema transacional do Grupo Marista, estão preenchidas para as disciplinas a qual cada ano possui (sendo nulas para caso não seja ofertada).

5.1.6 Características temporais

Devido ao desempenho do aluno variar com o ano, e a sua aprovação e satisfação final ser decorrente do progresso durante o ano, faz-se necessário o uso de características que descrevam esse comportamento temporal dos dados dos alunos.

5.1.6.1 Valores acumulados

Para demonstrar o progresso da vida acadêmica e *status* financeiro do aluno durante o ano, foram calculadas as medidas acumuladas, descritas pela Equação 4.3, durante o ano. As colunas adicionadas estão disponíveis na Tabela 5.6.

Tabela 5.6: Atributos temporais acumulados no conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
70	Faltas acumuladas	100,00%	*float*
71	$N_{acumulado}$ Artes	100,00%	*float*
72	$N_{acumulado}$ Biologia	100,00%	*float*
73	$N_{acumulado}$ Ciências	100,00%	*float*
74	$N_{acumulado}$ Filosofia	100,00%	*float*
75	$N_{acumulado}$ Física	100,00%	*float*
76	$N_{acumulado}$ Geografia	100,00%	*float*
77	$N_{acumulado}$ História	100,00%	*float*
78	$N_{acumulado}$ Inglês	100,00%	*float*
79	$N_{acumulado}$ Português	100,00%	*float*
80	$N_{acumulado}$ Química	100,00%	*float*
81	$N_{acumulado}$ Sociologia	100,00%	*float*
82	$N_{acumulado}$ Educação Física	100,00%	*float*
83	$N_{acumulado}$ Matemática	100,00%	*float*

5.1.6.2 Variação trimestral

Para avaliar quedas e melhorias de desempenho estudantil durante o decorrer do ano, foram adicionadas colunas de variação da nota, faltas e valores entre trimestres para os alunos. A Tabela 5.7 demonstra os dados presentes no conjunto de dados.

Tabela 5.7: Atributos temporais de variação no conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
84	$\Delta N_{trimestral}$ Faltas	100,00%	*float*
85	$\Delta N_{trimestral}$ Artes	100,00%	*float*
86	$\Delta N_{trimestral}$ Biologia	100,00%	*float*
87	$\Delta N_{trimestral}$ Ciências	100,00%	*float*
88	$\Delta N_{trimestral}$ Educação Física	100,00%	*float*

89	$\Delta N_{trimestral}$ Filosofia	100,00%	*float*
90	$\Delta N_{trimestral}$ Física	100,00%	*float*
91	$\Delta N_{trimestral}$ Geografia	100,00%	*float*
92	$\Delta N_{trimestral}$ História	100,00%	*float*
93	$\Delta N_{trimestral}$ Inglês	100,00%	*float*
94	$\Delta N_{trimestral}$ Matemática	100,00%	*float*
95	$\Delta N_{trimestral}$ Português	100,00%	*float*
96	$\Delta N_{trimestral}$ Química	100,00%	*float*
97	$\Delta N_{trimestral}$ Sociologia	100,00%	*float*

5.1.6.3 Variação anual

Para avaliar quedas e melhorias de desempenho estudantil entre diferentes anos, foi adicionado também a variação da nota, faltas e valores entre trimestres para os alunos. A Tabela 5.8 demonstra os dados presentes no conjunto de dados.

Tabela 5.8: Atributos temporais do conjunto de dados.

Número	**Nome da feature**	**Não-nulos**	**Tipo**
98	Variação anual de anuidade	100,00%	*float*
99	Variação anual de desconto	100,00%	*float*
100	Variação anual do total de faltas	100,00%	*float*
101	ΔN_{anual} Artes	100,00%	*float*
102	ΔN_{anual} Biologia	100,00%	*float*
103	ΔN_{anual} Ciências	100,00%	*float*
104	ΔN_{anual} Educação Física	100,00%	*float*
105	ΔN_{anual} Filosofia	100,00%	*float*
106	ΔN_{anual} Física	100,00%	*float*
107	ΔN_{anual} Geografia	100,00%	*float*
108	ΔN_{anual} História	100,00%	*float*
109	ΔN_{anual} Inglês	100,00%	*float*
110	ΔN_{anual} Matemática	100,00%	*float*
111	ΔN_{anual} Português	100,00%	*float*
112	ΔN_{anual} Química	100,00%	*float*
113	ΔN_{anual} Sociologia	100,00%	*float*

5.2 Resultados dos classificadores

Usando a base obtida a partir dos dados do Grupo Marista, foi possível treinar uma série de algoritmos classificadores tendo como alvo prever a evasão dos alunos. Para as classificações, foram testadas diversas configurações de classificadores. As métricas avaliadas para a comparação dos classificadores foram:

- Área sob a curva precisão-revocação (PR AUC);
- F1-score;
- Área sob a curva característica de operação do receptor (ROC AUC); e
- Acurácia.

5.2.1 Previsão de evasão geral

Primeiramente, foram treinados múltiplos classificadores sobre a base completa, isto é, sem segmentação por trimestre ou segmento estudantil. Esses resultados servirão como *baseline* para os demais, não tendo fim como resultado final devido a não levar em consideração características inerentes dos trimestres ou diferenças entre segmentos. A Tabela 5.9 apresenta os resultados obtidos a partir dos classificadores.

Tabela 5.9: Resultados dos classificadores no contexto geral.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	97,33%	42,64%	30,93%	92,87%
DecisionTreeClassifier	97,61%	36,77%	38,34%	84,74%
LogisticRegression	97,22%	32,80%	19,49%	87,50%
RandomForestClassifier	97,31%	44,44%	15,94%	90,37%
XGBClassifier	**98,42%**	**75,35%**	**66,73%**	**96,91%**
Média Geral	*97,58%*	*46,40%*	*34,29%*	*90,48%*

Para a classificação geral, pode-se notar que o algoritmo *XGBClassifier* obteve melhores resultados em todas as métricas, visto os maiores valores obtidos pelo classificador quando comparado aos demais. Em casos de classificação onde não existe discriminação, ou separação, entre exemplos de diferentes segmentos de ensino ou épocas do ano, ele é o melhor algoritmo dentre as abordagens propostas. O algoritmo acertou 98,42% dos casos de teste apresentados após o treino.

Devido a maior relevância da detecção de casos de evasão e o grande desequilí-

brio da base, a taxa verdadeiros negativos ou seja, classificados corretamente como não-evasores, é de menor relevância. Portanto, a avaliação das demais métricas é de grande importância.

O classificador apresentou um ROC AUC de 96,91%. Isso significa uma boa capacidade de distinção entre as classes evasor e não evasor, devido à razão de verdadeiros positivos com relação a falsos positivos ser alta.

O classificador apresentou F_1-*score* de 66,73%. Isso implica em uma qualidade do modelo superior aos demais, visto que a combinação de precisão e revocação é muito maior.

O classificador apresentou PR AUC de 75,35%. Isso implica que a capacidade de detectar casos positivos é vastamente superior aos outros algoritmos, característica extremamente desejada, pois é diretamente proporcional a capacidade do modelo de detectar casos positivos de evasão.

5.2.2 Previsão de evasão por segmento de ensino

Devido à grande diferença entre os perfis de estudante entre os segmentos de ensino, do infantil até o médio, é possível separar os alunos com intuito de obter melhores resultados em cada segmento.

5.2.2.1 Ensino Infantil

Após separados os alunos do Ensino Infantil do restante da base, os algoritmos foram treinados novamente sobre esta nova base, os quais tem seus resultados dispostos na Tabela 5.10.

Tabela 5.10: Resultados dos classificadores no Ensino Infantil.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	93,57%	64,38%	56,63%	91,46%
DecisionTreeClassifier	94,89%	69,80%	67,31%	91,24%
LogisticRegression	92,08%	49,79%	38,77%	86,86%
RandomForestClassifier	91,44%	67,25%	16,37%	91,04%
XGBClassifier	**95,78%**	**82,26%**	**73,74%**	**95,29%**
Média Geral	*93,55%*	*66,69%*	*50,56%*	*91,18%*

Para a classificação no Ensino Infatil, pode-se notar que o algoritmo *XGBClassifier*, novamente, obteve melhores resultados em todas as métricas relevantes à detecção de

evasão. O algoritmo obteve acurácia de 95.78% dos testes, sinalizando a maior taxa de acerto dentre os algoritmos.

Entretanto, os valores de ROC AUC e Acurácia são menores do que os apresentados no caso geral (Tabela 5.9), enquanto os valores de PR AUC são maiores. Sendo assim, pode-se afirmar que o modelo obteve melhores resultados para prever um caso positivo de evasão, ao custo de diminuir o número de verdadeiros casos de não evasão.

5.2.2.2 Ensino Fundamental - Anos Iniciais

Após separados os alunos do Ensino Fundamental, apenas para os Anos Iniciais, do restante da base, os algoritmos foram treinados novamente em cima desta nova base, os quais tem seus resultados dispostos na Tabela 5.11.

Tabela 5.11: Resultados dos classificadores no Ensino Fundamental - Anos Iniciais.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	98,28%	35,84%	27,86%	93,40%
DecisionTreeClassifier	98,40%	28,47%	30,64%	75,44%
LogisticRegression	98,18%	17,93%	5,43%	85,04%
RandomForestClassifier	98,28%	39,73%	8,33%	90,93%
XGBClassifier	**98,87%**	**66,84%**	**58,72%**	**96,64%**
Média Geral	*98,40%*	*37,76%*	*26,20%*	*88,29%*

No caso da previsão de evasão no Ensino Fundamental - Anos Iniciais, pode-se constatar que o *XGBClassifier* obteve novamente os melhores resultados, classificando corretamente 98,87% dos exemplos. Dito isso, os valores obtidos para todas as métricas relevantes à detecção de evasão foram, em média, menores do que o caso geral, conforme apresentado na Tabela 5.9.

5.2.2.3 Ensino Fundamental - Anos Finais

Após separados os alunos do Ensino Fundamental, apenas para os anos finais, do restante da base, os algoritmos foram treinados novamente em cima desta nova base, os quais tem seus resultados dispostos na Tabela 5.12.

Tabela 5.12: Resultados dos classificadores no Ensino Fundamental - Anos Finais.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	98,15%	39,93%	33,20%	93,04%

DecisionTreeClassifier	98,25%	31,56%	33,82%	74,08%
LogisticRegression	98,04%	27,12%	18,13%	85,82%
RandomForestClassifier	98,15%	42,06%	15,79%	90,87%
XGBClassifier	**98,80%**	**68,25%**	**61,62%**	**95,98%**
Média Geral	*98,28%*	*41,78%*	*32,51%*	*87,96%*

O segmento de Anos Finais do Ensino Fundamental apresentou um comportamento similar ao de Anos Iniciais, visto que pode-se constatar que o *XGBClassifier* obteve novamente os melhores resultados, classificando corretamente 98,80% dos exemplos e os valores obtidos para todas as métricas relevantes à detecção de evasão foram em média menores do que o caso geral.

Devido a isso, pode-se afirmar que existe maior dificuldade de classificar a evasão no Ensino Fundamental, ou a falta de melhores descritores para esse segmento com relação aos demais na base desenvolvida.

5.2.2.4 Ensino Médio

Após separados os alunos do Ensino Médio do restante da base, os algoritmos foram treinados novamente em cima desta nova base, os quais tem seus resultados dispostos na Tabela 5.13.

Tabela 5.13: Resultados dos classificadores no Ensino Médio.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	96,82%	59,47%	50,87%	94,07%
DecisionTreeClassifier	97,18%	54,13%	55,81%	88,08%
LogisticRegression	96,77%	55,29%	49,35%	92,12%
RandomForestClassifier	96,86%	62,40%	42,25%	93,66%
XGBClassifier	**98,00%**	**79,71%**	**71,66%**	**96,83%**
Média Geral	*97,13%*	*62,20%*	*53,99%*	*92,95%*

O segmento de Ensino Médio obteve como melhor classificador o *XGBClassifier*, o qual atingiu 98% de acurácia nos testes executados. O segmento apresentou também os melhores resultados nas demais métricas, as quais tem maior importância na detecção de evasão, do que todos os outros segmentos de ensino. Além disso, o segmento obteve resultados melhores que o caso geral. Sendo assim, pode-se afirmar que a base tem mais características adequadas a classificação desse segmento.

5.2.3 Cortes trimestrais

Devido à natureza temporal do ano letivo nas escolas do Grupo Marista, é de interesse separar o caso em diferentes trimestres para atuar nos diferentes momentos do ano. Isto se deve ao fato de existirem características sazonais atribuídas a cada trimestre do ano.

Embora o comportamento dos classificadores perante os segmentos estudantis sejam diferentes entre si, obtendo diferentes valores nas métricas avaliadas, os resultados quando avaliados ao passar dos trimestres apresentou comportamento similar em todos os segmentos: o desempenho dos classificadores fica melhor com o passar do ano letivo.

Sendo assim, os resultados dos segmentos estudantis foram agrupados por cada trimestre a fim de fornecer uma explicação geral simplificada do comportamento com o passar do tempo. Por sua vez, as informações referentes aos resultados não agrupados, com finalidade de validação de combinações específicas de segmentos e trimestres, está disponível no Apêndice A.

5.2.3.1 Primeiro trimestre

Após separar os dados relativos apenas ao primeiro trimestre do ano letivo, os algoritmos classificadores foram treinados nesta porção da base de cada segmento estudantil. Os resultados obtidos podem ser visualizados na Tabela 5.14.

Tabela 5.14: Resultados dos classificadores no primeiro trimestre.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	97,25%	40,68%	33,95%	89,00%
DecisionTreeClassifier	97,27%	30,91%	31,42%	76,30%
LogisticRegression	97,02%	29,94%	18,89%	82,49%
RandomForestClassifier	97,18%	37,99%	19,74%	85,05%
XGBClassifier	**97,90%**	**62,31%**	**55,34%**	**93,69%**
Média Geral	*97,32%*	*40,37%*	*31,87%*	*85,30%*

A classificação, tomando como posto de vista o primeiro trimestre do ano letivo, apresentou 97,9% dos casos classificados corretamente. Dito isso, os resultados aqui obtidos são piores do que o caso geral de classificação, apresentados na Tabela 5.9, ambos obtidos usando o algoritmo *XGBClassifier*. Tal constatação ilustra uma possível falta de informações durante o início do ano letivo para a previsão de evasão.

5.2.3.2 Segundo trimestre

Após separar os dados relativos apenas no segundo trimestre do ano letivo, os algoritmos classificadores foram treinados nesta porção da base de cada segmento estudantil. Os resultados obtidos podem ser visualizados na Tabela 5.15.

Tabela 5.15: Resultados dos classificadores no segundo trimestre.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	98,83%	72,53%	68,66%	95,24%
DecisionTreeClassifier	98,82%	55,40%	66,07%	79,65%
LogisticRegression	98,64%	62,94%	60,22%	90,39%
RandomForestClassifier	98,63%	70,95%	56,50%	93,31%
XGBClassifier	**99,19%**	**82,01%**	**78,93%**	**96,51%**
Média Geral	*98,82%*	*68,77%*	*66,08%*	*91,02%*

A classificação, quando tomada do ponto de vista do segundo trimestre, demonstrou significativa melhora com relação ao primeiro trimestre. O algoritmo *XGBClassifier* classificou corretamente 99,19% dos exemplos usados como teste. Além disso, apresentou resultados melhores que o caso geral, conforme apresentado na Tabela 5.9, e que o primeiro trimestre, conforme apresentado na Tabela 5.14. Isso implica em um maior sucesso em prever corretamente as classes dos exemplos nesse momento do ano, com relação ao primeiro trimestre.

Tomando como referência a PR AUC, o classificador apresentou 82,01% no segundo trimestre contra 62,31% obtidos no primeiro trimestre. Isso implica em uma capacidade de captar casos de evasão muito maior nesse momento do ano. Isso é importante devido a grande relevância de detectar esses casos positivos.

5.2.3.3 Terceiro trimestre

Após separar os dados relativos apenas ao terceiro trimestre do ano letivo, os algoritmos classificadores foram treinados nesta porção da base de cada segmento estudantil. Os resultados obtidos podem ser visualizados na Tabela 5.16.

Tabela 5.16: Resultados dos classificadores no terceiro trimestre.

	Acurácia	**PR AUC**	**F1-score**	**ROC AUC**
AdaBoostClassifier	99,52%	77,55%	72,10%	99,08%
DecisionTreeClassifier	99,46%	51,37%	64,96%	78,73%

LogisticRegression	99,26%	50,77%	47,89%	93,23%
RandomForestClassifier	99,34%	68,82%	49,22%	97,73%
XGBClassifier	**99,76%**	**92,33%**	**86,38%**	**99,46%**
Média Geral	*99,47%*	*68,17%*	*64,11%*	*93,65%*

A classificação, quando tomada do ponto de vista do terceiro trimestre, demonstrou significativa melhora com relação ao primeiro e segundo trimestres. O algoritmo *XGBClassifier* classificou corretamente 99,76% dos exemplos usados como teste. Além disso, apresentou resultados melhores que o segundo trimestre, conforme apresentado na Tabela 5.15. Isso implica em um maior sucesso em prever corretamente as classes dos exemplos nesse momento do ano, com relação ao segundo trimestre.

Levando em consideração a melhora entre trimestres, é possível afirmar que o problema de previsão de evasão se torna de menor complexidade com o passar do ano. Tal diferença provavelmente se deve a maior quantidade de dados dos alunos e ao uso dos atributos derivados, hipótese que poderá ser constatada ao avaliar as interpretações dos modelos.

5.3 Interpretação de modelos e resultados

Após o treino dos modelos, foi realizada a interpretação dos resultados obtidos por eles obtidos. Primeiramente, foi realizada análise das probabilidades de saída dos alunos evasores. Sinalizando quais os alunos com maior risco de evasão por meio de um alerta indicativo da gravidade esperada do caso. Em seguida, foram analisados os fatores de maior contribuição para evasão, e as interpretações por trás das classificações obtidas pelos modelos.

5.3.1 Sinalização de estudantes de risco

A partir dos dados obtidos, foi gerada uma lista de alunos e seus respectivos riscos de evasão em relação ao tempo, para diferentes níveis de risco. A Tabela 5.17 ilustra dados fictícios semelhantes às informações que foram fornecidas a rede de escolas do Grupo Marista.

Com base nas informações da tabela acima e nas interpretações geradas pelo SHAP, é possível também gerar visualizações e *dashboards* a fim de monitoramento dos alunos. A Figura 5.1 ilustra o *mockup* de sugestão de dashboard, com dados de um estudante fictício para monitorar potenciais evasores.

Tabela 5.17: Distribuição das classes por segmento de ensino.

Ano	Aluno	Escola	Segmento	Série	Alerta	Chance (%)
2019	123456	Escola 1	EM	2ºB	VERMELHO	25,78%
2019	654321	Escola 1	EM	1ºA	VERMELHO	71,84%
2019	987456	Escola 1	EM	3ºC	VERMELHO	37,47%
2019	456789	Escola 1	EM	1ºD	VERMELHO	10,54%

Figura 5.1: Sugestão de *dashboard* para monitoração de evasão.

Fonte: O autor

O *mockup* da sugestão apresentada informa os dados básicos do aluno como a escola e município de moradia. O painel conta com informações simplificadas da situação do aluno como o nível de gravidade do caso, os fatores de maior risco e um gráfico ilustrando a probabilidade de evasão. A visão simplificada, com poucos números e com a informação representada graficamente, é ideal para o diagnóstico rápido dos estudantes. Possibilitando a atuação sobre um grande número de alunos de maneira rápida e efetiva.

A implementação de um *dashboard* contendo as informações presentes no *mockup* não pertence ao escopo desse trabalho, a qual é tida como encargo do Grupo Marista.

5.3.2 Atributos de importância

Feita a análise a partir das probabilidades de evasão, é de interesse também entender os motivos que potencialmente levam a uma evasão ou que possuam correlação com esta ação. Primeiramente, foram analisados os atributos de maior impacto nos modelos de maneira gráfica e tabular. Um exemplo de gráfico extraído pode ser visto na Figura 5.2, ilustrando os atributos mais importantes para o modelo previsor de evasão para o Ensino Fundamental - Anos Finais, no segundo trimestre de 2018.

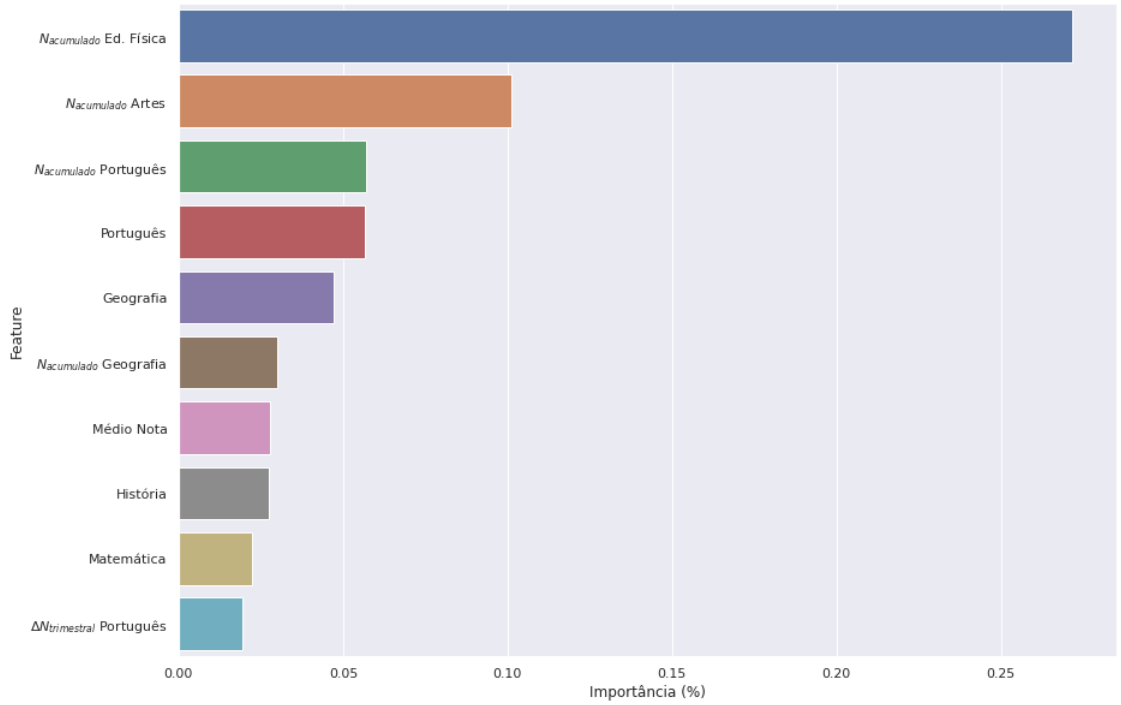

Figura 5.2: Exemplo de gráfico obtido a partir dos atributos de maior importância.

Fonte: O autor

Constatamos que a partir do gráfico, a forte relação entre as notas de Artes e Educação Física e a permanência do aluno na escola. Nota-se também que estas, como um todo, possuem um grande impacto no segmento estudantil em questão. Entretanto, o perfil apresentado por este segmento de ensino não é o mesmo quando comparado aos demais. As Tabelas 5.18, 5.19, 5.20 e 5.21 ilustram respectivamente os atributos que historicamente tem maior impacto na previsão de evasão nos segmentos Ensino Infantil, Ensino Fundamental - Anos Iniciais, Ensino Fundamental - Anos Finais e Ensino Médio.

Conforme apresentado pela Tabela 5.18, constatamos que notas não são um fator de impacto no Ensino Infantil. Visto que este segmento é não-obrigatório e tampouco contém forma unificada de avaliação dentre as escolas brasileiras. Notamos que os valores de anuidade e a quantidade de falta dos alunos são bons atributos para a previsão da

Tabela 5.18: Atributos mais importantes para os modelos previsores de evasão no Ensino Infantil.

Rank	**Segmento**	**Atributo**	**Importância**
1	EI	Anuidade	0,07
2	EI	Índice de Escolaridade do Município	0,06
3	EI	IDHMR	0,05
4	EI	Quantidade de Faltas	0,04
5	EI	ΔN_{anual} Anuidade	0,04

evasão neste segmento estudantil.

Verificamos também a aparição de atributos derivados ΔN e $N_{acumulado}$ dentre os mais importantes para previsão de evasão no segmento estudantil Ensino Fundamental - Anos Finais. Este fator sinaliza que os atributos adicionados ao conjunto de dados servem como melhores preditores do que as notas do trimestre, apenas.

Além disso, variáveis geográficas adicionadas ao conjunto de dados como o IDHMR e o Índice de Escolaridade Municipal, apresentaram-se como bons indicadores de evasão no Ensino Infantil, informando que dados de indicadores socioeconômicos podem impactar diretamente a probabilidade de evasão de um aluno. Entretanto, este retrato não é o mesmo quando comparado ao do Ensino Fundamental - Anos Iniciais, conforme demonstrado pela Tabela 5.19.

Tabela 5.19: Atributos mais importantes para os modelos previsores de evasão no Ensino Fundamental - Anos Iniciais.

Rank	**Segmento**	**Atributo**	**Importância**
1	EFAI	$N_{acumulado}$ Ed. Física	0,12
2	EFAI	$N_{acumulado}$ Artes	0,07
3	EFAI	Média das Notas	0,05
4	EFAI	$N_{acumulado}$ Inglês	0,04
5	EFAI	$N_{acumulado}$ Português	0,02

A Tabela 5.19 indica que o retrato nos Anos Iniciais do Ensino Fundamental é diferente do Ensino Infantil. Nota-se que as disciplinas normalmente tidas como mais lúdicas, como Educação Física e Artes, juntamente com as disciplinas voltadas à alfabetização do aluno como Inglês e Português, tem grande relação com a permanência do aluno na escola.

Verificamos que boa parte dos atributos mais importantes para os modelo previsores de evasão no Ensino Fundamental - Anos Iniciais advém da metodologia proposta para criação de atributos. As variáveis acumuladas das notas $N_{acumulado}$, descritas pela fórmula 4.3 apareceram em 80% das cinco variáveis de maior impacto nos modelos previsores de evasão nos Anos Iniciais.

O retrato apresentado pelos Ensino Fundamental - Anos Iniciais apresenta semelhanças com o segmento do Ensino Fundamental - Anos Finais, representado pela Tabela 5.20 e pela Figura 5.2.

Tabela 5.20: Atributos mais importantes para os modelos previsores de evasão no Ensino Fundamental - Anos Finais.

Rank	**Segmento**	**Atributo**	**Importância**
1	EFAF	$N_{acumulado}$ Ed. Física	0,13
2	EFAF	$N_{acumulado}$ Artes	0,06
3	EFAF	Média das Notas	0,05
4	EFAF	Português	0,04
5	EFAF	Geografia	0,04

Semelhantemente aos Anos Iniciais do Ensino Fundamental, a previsão de evasão nos Anos Finais é altamente impactada também pela nota das disciplinas como Artes e Educação Física. Embora a "Média das Notas" e a nota na disciplina de Português também serem fatores relevantes nos modelos previsores para os Anos Finais, a disciplina de Geografia mostra ser mais importante para a classificação de evasão neste segmento quando comparada aos Anos Iniciais.

Dentre os atributos de maior importância existem duas ocorrências de atributos derivados $N_{acumulado}$, sinalizando novamente que estes atributos servem como preditores relevantes para a evasão estudantil. Ponto que também acontece no Ensino Médio, conforme a Tabela 5.21.

Tabela 5.21: Atributos mais importantes para os modelos previsores de evasão no Ensino Médio.

Rank	**Segmento**	**Atributo**	**Importância**
1	EM	$N_{acumulado}$ Português	0,11
2	EM	$N_{acumulado}$ Geografia	0,09
3	EM	Média das Notas	0,06
4	EM	Português	0,03
5	EM	Menor Nota	0,03

A presença dos atributos derivados $N_{acumulado}$ entre os mais importantes também acontece no Ensino Médio, onde a performance acumulada das disciplinas de Português e Geografia tem alta relação com a permanência de um estudante. Notamos que as disciplinas de Artes e Educação Física não tem grande correlação na evasão do Ensino Médio, dando lugar ao valor "Menor Nota" do aluno e a nota de Geografia. Essa mudança pode ser explicada tanto por uma menor carga horária neste segmento de ensino ou ainda por menor interesse por parte dos alunos.

5.3.3 Interpretação dos modelos

A partir dos modelos treinados também é possível obter interpretações utilizando o SHAP. Com o auxílio deste conjunto de algoritmos, foram obtidos diferentes gráficos contemplando a base completa, assim como gráficos contemplando estudantes individuais.

Serão adotados os alunos do Ensino Fundamental - Anos Finais (EFAF) como ilustração dos gráficos obtidos. Isso se dá pelo fato do alto número de modelos obtidos, visto que são geradas todas as combinações possíveis entre segmento estudantil, trimestre e ano letivo.

A Figura 5.3 demonstra o *Decision Plot* obtido, ilustrando quais atributos levaram diversos estudantes a serem marcados como evasores. Por meio do gráfico disponibilizado pelo pacote SHAP.

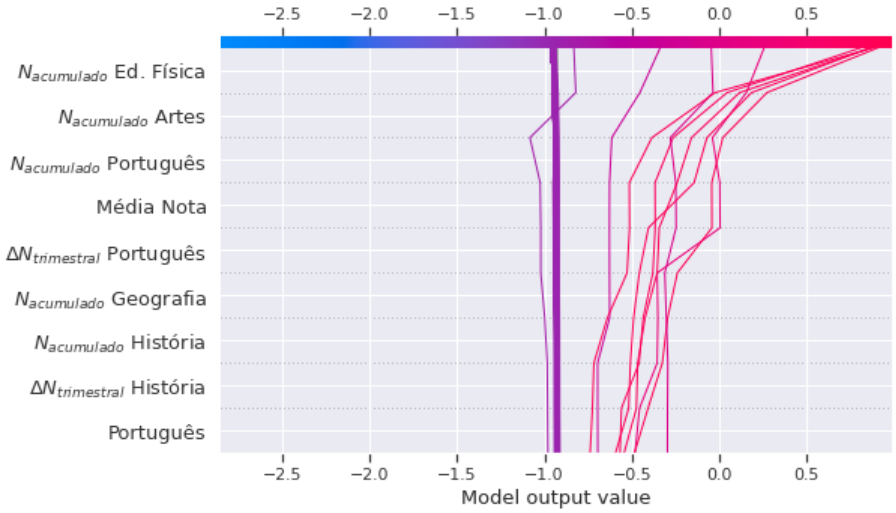

Figura 5.3: *Decision plot* obtido para uma amostra de estudantes do EFAF.

Fonte: O autor

Cada linha no *Decision Plot* da Figura 5.3 representa um exemplo apresentado ao classificador e seu respectivo desvio da média dos resultados. Caso um atributo tenha impacto na ocorrência da classe positiva (neste caso a evasão estudantil), ele será mostrado desviando o valor da trajetória para a direita, no sentido dos valores positivos. A probabilidade de um aluno ser marcado como evasor é mais alta quanto mais a direita a sua trajetória termina. O gráfico representa os atributos que o algoritmo do SHAP marcou como mais impactantes na parte superior do gráfico.

Observamos também que a maioria dos casos dos alunos com alta probabilidade de evasão no Ensino Fundamental - Anos Finais apresenta como fatores de maior impacto, as suas notas acumuladas $N_{acumulado}$ em Educação Física, Artes e Português, dados que

confirmam os valores de importância anteriormente observados na Tabela 5.20.

Embora a interpretação de um gráfico geral seja de interesse para a avaliação do panorama geral da situação dos estudantes, é de grande interesse extrair interpretações de classificações específicas para melhor compreender as causas de uma evasão.

Com o auxílio do SHAP, pode-se também tratar de casos individuais de evasão, possibilitando análise de por exemplo, casos verdadeiro positivos, falso positivos, falso negativos e verdadeiro negativos. As Figuras 5.4, 5.5, 5.6 e 5.7 ilustram respectivamente exemplos de estudantes em cada uma dessas categorias.

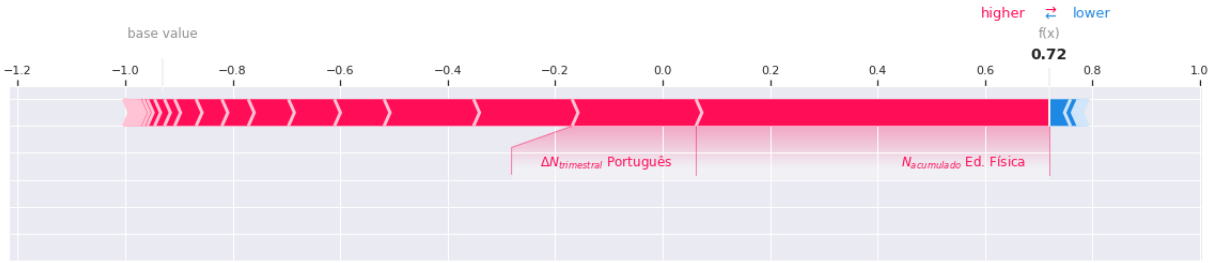

Figura 5.4: Atributos impactando um caso individual de verdadeiro positivo.

Fonte: O autor

A Figura 5.4 ilustra o caso de um aluno evasor, ilustrando os fatores de causa da sua classificação. Atributos que contribuem com a evasão são apresentados em vermelho, direcionando o valor de $f(x)$ para direita, enquanto os que contribuem com a permanência são representados pela cor azul e direcionam o valor de $f(x)$ para a esquerda.

O valor de $f(x)$ é usado pelo o SHAP para sinalizar o quão forte é o desvio do exemplo apresentado em direção à classe positiva. Sendo assim, é possível notar que o valor desta função é de 0.72, este sendo alto quando comparado com o *base value* do modelo.

Os motivos prováveis que levaram a saída do aluno foram a sua baixa nota acumulada $N_{acumulado}$ em Educação Física, além da queda brusca de desempenho na disciplina de Português, quando comparados ao trimestre anterior. Seria possível então, tomar medidas preventivas para que a evasão não ocorresse.

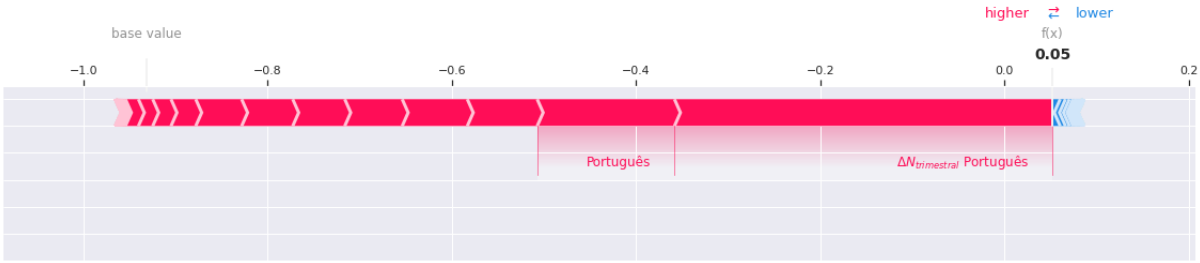

Figura 5.5: Atributos impactando um caso individual de falso positivo.

Fonte: O autor

A Figura 5.4 ilustra um caso de falso positivo por parte do modelo responsável pelo aluno em questão. É possível notar que o seu valor de $f(x)$ é de 0.05, alto quando comparado com o baixo *base value* do modelo. Isso é um indicador de que embora o classificador apresente esse aluno como evasor, não há tanta certeza da decisão tomada.

Não houve evasão por parte do aluno em questão, entretanto a existência de motivos que costumam ter alta correlação em casos de estudantes evasores, como sua nota baixa em português quando comparada ao trimestre anterior, levou ao erro do modelo. Embora esse aluno não tenha de fato evadido, um problema educacional como este poderia ser identificado facilmente, e medidas para melhorar a situação poderiam ter sido tomadas como por exemplo um reforço na disciplina.

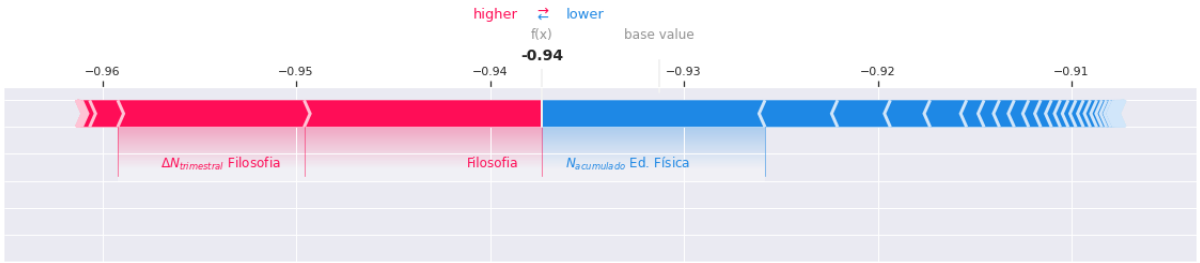

Figura 5.6: Atributos impactando um caso individual de falso negativo.

Fonte: O autor

A Figura 5.6 ilustra um caso onde o classificador falhou em identificar um aluno evasor. A partir do valor de $f(x)$ pode-se constatar que o algoritmo indica que o aluno era um retentor com alto grau de certeza. Isso é devido a sua alta nota acumulada em Educação Física em combinação com diversos outros fatores. Entretanto, a sua baixa nota em Filosofia comparada ao trimestre anterior é um potencial motivador de sua saída. Neste caso, embora o aluno não tenha sido categorizado como evasor, potenciais problemas na disciplina de Filosofia poderiam ter sido detectados.

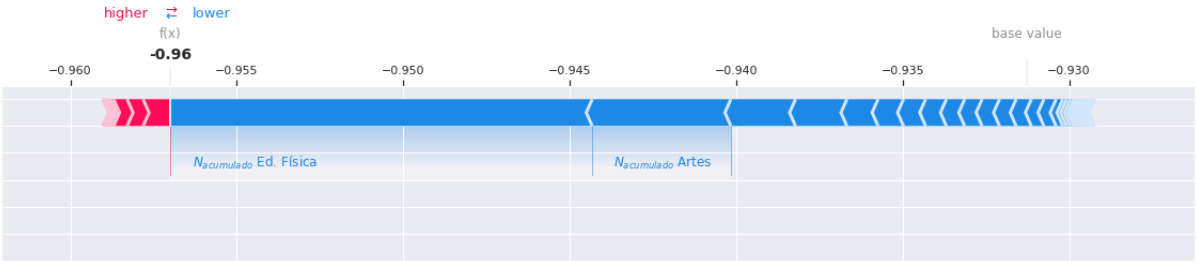

Figura 5.7: Atributos impactando um caso individual de verdadeiro negativo.

Fonte: O autor

Por fim, a Figura 5.7 ilustra um caso onde o classificador corretamente classificou um aluno como retentor. Observa-se que o classificador afirmou com alto grau de confiança a permanência do aluno. Isso é devido ao seu bom desempenho em Educação Física e Artes.

5.4 Discussão geral dos resultados

O processo de geração e tratamento do conjunto de dados foi desenvolvido com sucesso, o que pode ser constatado pelo desempenho dos algoritmos classificadores e análises exploratórias.

Após realizada a classificação nas diferentes repartições da base, pode-se observar alguns detalhes e pontos importantes a serem notados nos resultados obtidos.

Dentre os classificadores, o classificador por *Extreme Gradient Boosting Trees*, ou *XGBClassifier*, apresentou os melhores resultados em todas as repartições propostas da base original. O classificador também apresentou os melhores resultados na base completa, dado os valores maiores encontrados nas métricas avaliadas, na classificação da base completa. Este fato constata que a hipótese de que modelos caixa-branca são tão competitivos quando comparados a modelos caixa-preta na tarefa de previsão de evasão é falsa, sendo assim, necessário utilizar modelos caixa preta para esta tarefa.

O *XGBClassifier* apresentar melhores resultados quando comparado aos demais algoritmos se deve provavelmente a estrutura gerada pelo algoritmo. O treino de classificadores dentro do *ensemble*, os quais recebem uma parcela dos atributos disponíveis, viabiliza a detecção de interações entre múltiplas variáveis. Interações as quais podem ter alto grau de importância ao se prever casos de evasão.

5.4.1 Evasão nos segmentos de ensino

Levando em consideração os diferentes segmentos de ensino e os resultados obtidos, é possível constatar que existem diferenças entre os diferentes segmentos de ensino, conforme anteriormente suposto. Além disso, pode-se notar que os segmentos que apresentaram melhores resultados que os demais foram os Ensino Infantil e Ensino Médio, com o Ensino Médio se destacando com o com melhores resultados dentre os dois.

Os segmentos respectivos ao Ensino Fundamental apresentaram piores resultados que os demais, com destaque para os Anos Iniciais, onde os valores das métricas avaliadas apresentaram-se como os menores. Isso pode indicar que é um segmento onde a previsão é mais difícil ou ainda que não existem bons descritores para este segmento no conjunto de dados.

5.4.2 Evasão nos trimestres

Levando em consideração os diferentes momentos do ano, houve uma melhoria expressiva do primeiro trimestre para o segundo e terceiro, nos quais os modelos possuem performance similar. Sendo assim, um bom momento para classificar e abordar os potenciais evasores é a partir do segundo trimestre, visto que ainda há tempo para atuar na retenção do aluno e os modelos apresentam bons resultados.

Além disso, foi possível constatar que, ao decorrer do ano, a tarefa de classificação fica mais fácil, conforme inicialmente suposto. Isso se deve, à maior quantidade de dados coletados referentes ao ano letivo e ao maior impacto dos atributos derivados ΔN e $N_{acumulado}$.

5.4.3 Interpretação dos modelos

A partir dos resultados obtidos através das importâncias dos atributos e dos gráficos do SHAP é possível realizar diversas constatações relacionadas às *features* levantadas e aos segmentos estudantis com base no chamado *XGBClassifier* (CHEN; GUESTRIN, 2016), o melhor modelo encontrado para previsão de evasão.

A hipótese inicial de que dados financeiros, socioeconômicos e dados relativos a localização da moradia do aluno, podem auxiliar na previsão de casos de evasão foi constatada por meio das importâncias destes atributos nos modelos previsores de evasão no Ensino Infantil. Atributos como IDHMR e o Índice de Escolaridade do Município são bons indicadores de evasão neste segmento de ensino. Além disso, notamos também que dados financeiros também tem impacto na previsão neste segmento estudantil, com o valor da

anuidade sendo o maior indicador de um aluno sair da escola.

Diferente do Ensino Infantil, os segmentos Ensino Fundamental - Anos Iniciais e Anos Finais tiveram as notas como maiores fatores indicadores de evasão. A performance no decorrer do ano, dada pelo atributo derivado $N_{acumulado}$, nas disciplinas de Artes e Educação Física, se mostrou um bom indicador para prever a evasão de um aluno. Pode-se notar também que as notas nas disciplinas de línguas, como Inglês e Português, também apresentaram um impacto na determinação da permanência dos alunos nas escolas. Fato demonstrado tanto em gráficos de casos individuais quanto no retrato geral.

Em contraste ao Ensino Fundamental, os modelos previsores de evasão para alunos do Ensino Médio dão maior importância para disciplinas como Português e Geografia. Entretanto, as notas do aluno ainda são os maiores fatores de impacto na tarefa de previsão de sua permanência ou saída da escola.

Embora o uso das importâncias fornecidas pelos algoritmos e das interpretações fornecidas pelo SHAP indiquem a correlação entre a variável alvo e os atributos acima elencados, é importante notar que essa correlação não implica necessariamente em uma relação de causalidade. Devido a complexa natureza do problema (o qual envolve questões financeiras, pessoais e sociais) e a dificuldade de modelar completamente a situação do aluno, não é possível afirmar indubitavelmente que os motivos apontados sejam de fatos os causadores de evasões.

Por sua vez, a construção de hipóteses é possível a partir das interpretações geradas pelo conjunto de técnicas apresentado. Possibilitando a sugestão de medidas preventivas para a evasão. As construções geradas abordam o objetivo de gerar interpretações a partir de modelos caixa-preta no contexto da evasão estudantil. Construções estas que podem ser usadas a fim de um melhor diagnóstico para o problema de evasão nas escolas.

Capítulo 6

Conclusão

O problema da evasão estudantil é de grande impacto social, afetando as futuras vidas de crianças e adolescentes ao redor do mundo, visto que são o futuro do mercado de trabalho. Se por um lado impacta socialmente, por outro também impacta o mercado da educação, este sendo altamente competitivo, que também movimenta e emprega milhões de pessoas. Uma maior retenção significa um maior lucro, e por sua vez, um maior retorno de investimento.

Houveram tentativas de prever a evasão de alunos, mas há grande dificuldade devido a variedade de diferentes contextos para se abordar. O grande desequilíbrio entre as classes e a dificuldade para se obter dados externos são fatores que tornam o problema ainda mais desafiador.

Um fator que auxilia na previsão de evasão é sua similaridade com o problema de previsão de *churn*, o qual apresenta características semelhantes ao ponto de vista de modelo de negócio. Isso possibilita o uso de inovações em um dado contexto e verificar se as mesmas são aplicáveis em outro.

Os modelos que apresentam os melhores resultados na previsão de evasão e na previsão de *churn* de clientes nem sempre são os interpretáveis para as equipes de negócio. Como exemplo, existem o uso de *ensembles* ou SVM para a classificação, os quais são exemplos de modelo caixa preta. Sendo assim, o uso de técnicas de explicação de modelos se mostra uma ferramenta útil para a formulação e comprovação de hipóteses.

Neste trabalho foi apresentado o processo de criação de uma base de dados contendo dados históricos dos alunos do Grupo Marista, desde a Educação Infantil até o Ensino Médio. Foram também apresentados os resultados dos classificadores prevendo a evasão de tais estudantes, no qual o *XGBClassifier* apresentou os melhores resultados com relação aos demais. Isso constatou que a hipótese inicial de que modelos caixa-branca são competitivos com modelos caixa-preta para a tarefa de previsão de evasão é falsa, e sendo

assim, é necessário utilizar modelos mais complexos.

As soluções levantadas para o problema de evasão apresentaram bons resultados, encontrando valores similares aos da literatura em uma base de dados mais abrangente e de criação simplificada.

Constatamos durante o decorrer do trabalho a hipótese de que os fatores socioeconômicos, como o IDH e Índice de Escolaridade, em conjunto das notas do aluno, foram fatores que impactaram positivamente na previsão de evasão na Educação Básica.

A partir dos modelos gerados foi possível obter fatores de maior impacto para os classificadores e interpretações para casos gerais e individuais de evasão. O SHAP (LUNDBERG; LEE, 2017) mostrou-se uma ferramenta adequada para a demonstração dos pontos de maior impacto numa evasão. Além de prover gráficos para melhor compreensão da interpretação por ele gerada. De tal modo foi possível atingir o objetivo de obter interpretações e explicações a partir de modelos caixa-preta, as quais normalmente são disponíveis apenas em modelos caixa-branca, no contexto da evasão estudantil.

Esperamos que com base nos resultados por este trabalho obtidos, seja possível diagnosticar alunos evasores e melhorar a taxa de retenção nas escolas. Fator que contribui para aumento da qualidade da Educação Básica, auxiliando no processo de criação de uma sociedade mais justa.

Como futuros estudos é proposto o acompanhamento do desempenho das técnicas apresentadas por este trabalho no decorrer do tempo e no segmento do Ensino Superior. Além disso, é possível também explorar o uso de técnicas para lidar com bases desbalanceadas, como o SMOTE (CHAWLA et al., 2002), para contornar o fator do desbalanceamento em bases de dados de evasão estudantil.

Referências Bibliográficas

ADELMAN, M. A.; SZEKELY, M. School dropout in central america: An overview of trends, causes, consequences, and promising interventions. *World Bank Policy Research Working Paper*, n. 7561, 2016.

AHMED, M.; AFZAL, H.; MAJEED, A.; KHAN, B. A survey of evolution in predictive models and impacting factors in customer churn. *Advances in Data Science and Adaptive Analysis*, World Scientific, v. 9, n. 03, p. 1750007, 2017.

AHN, J.; HWANG, J.; KIM, D.; CHOI, H.; KANG, S. A survey on churn analysis in various business domains. *IEEE Access*, IEEE, v. 8, p. 220816–220839, 2020.

ALMANA, A. M.; AKSOY, M. S.; ALZAHRANI, R. A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, v. 4, n. 5, p. 165–171, 2014.

AMAZON. *O que é boosting?* 2022. Disponível em: <https://aws.amazon.com/pt/what-is/boosting/>.

BACH, S.; BINDER, A.; MONTAVON, G.; KLAUSCHEN, F.; MÜLLER, K.-R.; SAMEK, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 7, p. e0130140, 2015.

BANDARA, W. M. C.; PERERA, A. S.; ALAHAKOON, D. Churn prediction methodologies in the telecommunications sector: A survey. In: *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*. [S.l.: s.n.], 2013. p. 172–176.

BARDACH, L.; LÜFTENEGGER, M.; OCZLON, S.; SPIEL, C.; SCHOBER, B. Context-related problems and university students' dropout intentions—the buffering effect of personal best goals. *European Journal of Psychology of Education*, v. 35, 08 2019.

BATISTA, S. D.; SOUZA, A. M.; OLIVEIRA, J. M. d. S. A evasão escolar no ensino médio: um estudo de caso. *Revista Profissão Docente, UNIUBE. Uberaba/MG*, v. 9, n. 19, 2009.

BEAL, P. E.; NOEL, L. What works in student retention: The report of a joint project of the american college testing program and the national center for higher education management systems. ERIC, 1980.

BERKSON, J. Application of the logistic function to bio-assay. *Journal of the American statistical association*, Taylor & Francis, v. 39, n. 227, p. 357–365, 1944.

BERSON, A.; THEARLING, K. *Building data mining applications for CRM*. [S.l.]: McGraw-Hill, Inc., 1999.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996, estabelece as diretrizes e bases da educação nacional. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 1996. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9394compilado.htm>.

BRASIL, U. *Enfrentamento da Cultura do Fracasso Escolar*. 2021. Disponível em: <https://www.unicef.org/brazil/relatorios/enfrentamento-da-cultura-do-fracasso-escolar>.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CERATTI, M. R. N. Evasão escolar: causas e consequências. *Curitiba/PR*, 2008.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

CORREIOS. *API dos correios*. Correios API, 2021. Disponível em: <https://cws.correios.com.br>.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

DAHIYA, K.; BHATIA, S. Customer churn analysis in telecom industry. In: IEEE. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO, Trends and Future Directions)*. [S.l.], 2015. p. 1–6.

DUMITRACHE, A.; NASTU, A. A.; STANCU, S. Churn prediction in telecommunication industry: Model interpretability. 2020.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação por escrito*, v. 8, n. 1, p. 35–48, 2017.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997.

FUNG, P. L.; ZAIDAN, M. A.; TIMONEN, H.; NIEMI, J. V.; KOUSA, A.; KUULA, J.; LUOMA, K.; TARKOMA, S.; PETäJä, T.; KULMALA, M.; HUSSEIN, T. Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. *Journal of Aerosol Science*, v. 152, p. 105694, 2021. ISSN 0021-8502. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021850220301798>.

FUSI, N.; SHETH, R.; ELIBOL, M. H. Probabilistic matrix factorization for automated machine learning. In: *NIPS 2018*. arXiv:1705.-5355v2, 2018. Preprint posted to Cornell University Library. Disponível em: <https://www.microsoft.com/en-us/research/publication/probabilistic-matrix-factorization-for-automated-machine-learning/>.

HABLEY, W. R.; MCCLANAHAN, R. What works in student retention. *Iowa City, IA: American College Testing Service. Retrieved February*, v. 21, p. 2005, 2004.

HEFLIN, H.; SHEWMAKER, J.; NGUYEN, J. Impact of mobile technology on student attitudes, engagement, and learning. *Computers and Education*, v. 107, p. 91–99, 2017. ISSN 0360-1315. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360131517300064>.

HUANG, B.; KECHADI, M. T.; BUCKLEY, B. Customer churn prediction in telecommunications. *Expert Systems with Applications*, Elsevier, v. 39, n. 1, p. 1414–1425, 2012.

IBGE. *Indicadores IBGE*. Pesquisas IBGE, 2021. Disponível em: <https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas.html>.

JADHAV, A.; PRAMOD, D.; RAMANATHAN, K. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, Taylor & Francis,

v. 33, n. 10, p. 913–933, 2019. Disponível em: <https://doi.org/10.1080/08839514.2019.1637138>.

JAFARI, R.; DENTON, J.; IDRIS, A.; SMITH, B.; KERAMATI, A. Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry. *Neural Computing and Applications*, 09 2020.

JENKINS, J. M.; DUNCAN, G. J.; AUGER, A.; BITLER, M.; DOMINA, T.; BURCHINAL, M. Boosting school readiness: Should preschool teachers target skills or the whole child? *Economics of Education Review*, v. 65, p. 107–125, 2018. ISSN 0272-7757. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0272775717302509>.

KAMALRAJ, N.; MALATHI, A. A survey on churn prediction techniques in communication sector. *International Journal of Computer Applications*, Citeseer, v. 64, n. 5, p. 39–42, 2013.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KAYONDA, H. N.; LOMBO, L. S.; LOMBO, G.; VIVIAR, N. E. Causes and consequences of school dropout in kinshasa: Students' perspectives before and after dropping out. *Journal of African Education*, Adonis & Abbey Publishers Ltd, v. 2, n. 3, p. 177, 2021.

KERR, P. Adaptive learning. *ELT Journal*, v. 70, n. 1, p. 88–93, 10 2015. ISSN 0951-0893. Disponível em: <https://doi.org/10.1093/elt/ccv055>.

KOMENAR, M. *Electronic marketing*. [S.l.]: Wiley Computer, 1997.

KUBAT, M.; MATWIN, S. et al. Addressing the curse of imbalanced training sets: one-sided selection. In: CITESEER. *Icml*. [S.l.], 1997. v. 97, p. 179–186.

LAU, L. K. Institutional factors affecting student retention. *Education*, v. 124, n. 1, 2003.

LEJEUNE, M. A. Measuring the impact of data mining on churn management. *Internet Research*, MCB UP Ltd, 2001.

LEON, F. L. L. d.; MENEZES-FILHO, N. A. Reprovação, avanço e evasão escolar no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2002.

LIN, X.; YACOUB, S.; BURNS, J.; SIMSKE, S. Performance analysis of pattern classifier combination by plurality voting. *Pattern Recognition Letters*, Elsevier, v. 24, n. 12, p. 1959–1969, 2003.

LINCOFF, G. H.; SCHLIMMER, J. *Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms*. [S.l.]: Alfred A. Knopf, 1981.

LOYOLA-GONZALEZ, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, IEEE, v. 7, p. 154096–154113, 2019.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

LUNDBERG, S. M.; NAIR, B.; VAVILALA, M. S.; HORIBE, M.; EISSES, M. J.; ADAMS, T.; LISTON, D. E.; LOW, D. K.-W.; NEWMAN, S.-F.; KIM, J. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, Nature Publishing Group, v. 2, n. 10, p. 749, 2018.

LYKOURENTZOU, I.; GIANNOUKOS, I.; NIKOLOPOULOS, V.; MPARDIS, G.; LOUMOS, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, v. 53, n. 3, p. 950–965, 2009. ISSN 0360-1315. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360131509001249>.

MANNESOFT. *Prime*. Mannesoft, 2021. Disponível em: <https://www.mannesoftprime.com.br/>.

MÁRQUEZ, C.; CANO, A.; ROMERO, C.; MOHAMMAD, A.; FARDOUN, H.; VENTURA, S. Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, v. 33, p. 107–124, 02 2016.

MEC. *Resumo técnico do Censo Educacional de 2019*. Ministério da Educação, 2019. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_basica_2019.pdf>.

MITCHELL, D.; COLES, C. Establishing a continuing business model innovation process. *Journal of Business Strategy*, v. 25, p. 39–49, 06 2004.

MIWA, S.; SHIMOSEGAWA, M.; HONGO, K.; SHINDO, M.; OBUCHI, M.; MATSUMOTO, E.; ARIMOTO, M.; CLARK, I.; YAMAMOTO, S.; SHINKAWA, M. Dropout

from higher education and social stratification in japan. *Annu Bull Graduate Sch Educ Tohoku Univ*, v. 1, p. 1–18, 2015.

MOLNAR, C. *Interpretable machine learning*. [S.l.]: Lulu. com, 2020.

NA, C. The consequences of school dropout among serious adolescent offenders: More offending? more arrest? both? *Journal of Research in Crime and Delinquency*, v. 54, n. 1, p. 78–110, 2017. Disponível em: <https://doi.org/10.1177/0022427816664118>.

O'KEEFFE, P. A sense of belonging: Improving student retention. *College Student Journal*, Project Innovation, v. 47, n. 4, p. 605–613, 2013.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PLATT, J. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*, v. 208, 07 1998.

QI, J.; ZHANG, Y.; SHU, H.; LI, Y.; GE, L. Churn prediction with limited information in fixed-line telecommunication. In: *Symposium on Communication Systems Networks and Digital Signal Processing*. [S.l.: s.n.], 2006. p. 423–426.

QUARTI, E.; FIGINI, S.; GIUDICI, P. Churn risk mitigation models for student's behavior. *Electronic Journal of Applied Statistical Analysis*, v. 2, 01 2009.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014.

RA, S.; SHRESTHA, U.; KHATIWADA, S.; YOON, S. W.; KWON, K. The rise of technology and impact on skills. *International Journal of Training Research*, Routledge, v. 17, n. sup1, p. 26–40, 2019. Disponível em: <https://doi.org/10.1080/14480220.2019.1629727>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence A Modern Approach*. [S.l.: s.n.], 2010.

SALES, A.; BALBY, L.; CAJUEIRO, A. Exploiting academic records for predicting student drop out: A case study in brazilian higher education. *Journal of Information and Data Management*, v. 7, n. 2, p. 166–166, 2016.

SCHINDLER, L. A.; BURKHOLDER, G. J.; MORAD, O. A.; MARSH, C. Computer-based technology and student engagement: a critical review of the literature. *International Journal of Educational Technology in Higher Education*, SpringerOpen, v. 14, n. 1, p. 1–28, 2017.

SHAPLEY, L. S. *Notes on the N-Person Game - II: The Value of an N-Person Game*. Santa Monica, CA: RAND Corporation, 1951.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2017. p. 3145–3153.

SOUSA, C. R. d. O.; GOMES, K. R. O.; SILVA, K. C. d. O.; MASCARENHAS, M. D. M.; RODRIGUES, M. T. P.; ANDRADE, J. X.; LEAL, M. A. B. F. Fatores preditores da evasão escolar entre adolescentes com experiência de gravidez. *Cadernos Saúde Coletiva*, SciELO Brasil, v. 26, p. 160–169, 2018.

STROUSE, K. G. *Marketing telecommunications services: new approaches for a changing environment*. [S.l.]: Artech House Publishers, 1999.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, Springer, v. 41, n. 3, p. 647–665, 2014.

TINTO, V. Research and practice of student retention: What next? *Journal of college student retention: Research, Theory & Practice*, SAGE Publications Sage CA: Los Angeles, CA, v. 8, n. 1, p. 1–19, 2006.

USA. Equal credit opportunity act - 15 u.s.c. §§ 1691-1691f. *Consumer Credit Protection Act*, 1974. Disponível em: <https://www.ftc.gov/enforcement/statutes/equal-credit-opportunity-act>.

VILLARREAL, D.; ZHAO, L.; HILL, T.; CHUNG, L. Validating goal-oriented hypotheses of business problems using machine learning: An exploratory study of customer churn. In: SPRINGER NATURE. *Big Data–BigData 2020: 9th International Conference, Held as Part of the Services Conference Federation, SCF 2020, Honolulu, HI, USA, September 18-20, 2020, Proceedings*. [S.l.], 2020. v. 12402, p. 144.

WESTERA, W. Reframing the role of educational media technologies. 2015.

ZAHARIA, M.; XIN, R. S.; WENDELL, P.; DAS, T.; ARMBRUST, M.; DAVE, A.; MENG, X.; ROSEN, J.; VENKATARAMAN, S.; FRANKLIN, M. J.; GHODSI, A.; GONZALEZ, J.; SHENKER, S.; STOICA, I. Apache spark: A unified engine for big data processing. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 59, n. 11, p. 56–65, out. 2016. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/2934664>.

Apêndice A

Resultados obtidos pelo classificador

Tabela A.1: Resultados dos classificadores no contexto geral

Segmento	**Algoritmo**	**N**	**Acurácia**	**PR AUC**	**F1-score**	**ROC AUC**
EI	XGBoost	NA	95,80%	18,08%	26,54%	57,78%
EFAI	XGBoost	1	98,31%	19,15%	32,87%	60,13%
EFAI	XGBoost	2	99,59%	67,40%	81,01%	85,41%
EFAI	XGBoost	3	99,91%	75,85%	86,53%	91,80%
EFAF	XGBoost	1	97,78%	23,45%	37,50%	61,81%
EFAF	XGBoost	2	99,31%	65,62%	79,21%	83,49%
EFAF	XGBoost	3	99,88%	82,31%	90,55%	93,77%
EM	XGBoost	1	95,67%	31,34%	45,42%	65,16%
EM	XGBoost	2	98,74%	74,00%	84,71%	87,73%
EM	XGBoost	3	99,65%	81,40%	90,02%	93,63%
Média			*98,46%*	*53,86%*	*65,44%*	*78,07%*

Apêndice B

Distribuição das colunas por segmento

Tabela B.1: Distribuição das colunas por segmento

Nome da Coluna	**Segmento**	**Distribuição**[1]	$\bar{x}$	σ	**Não-nulos (%)**
Anuidade	EI	Não Gaussiana	10694,33	7156,45	100
Anuidade	EFAI	Gaussiana	12444,84	7563,24	93,31
Anuidade	EFAF	Gaussiana	13522,79	8491,31	92,79
Anuidade	EM	Gaussiana	15592,52	10304,80	92,12
Ano Entrada	EI	Não Gaussiana	2016,36	2,05	100
Ano Entrada	EFAI	Não Gaussiana	2014,25	2,99	100
Ano Entrada	EFAF	Não Gaussiana	2012,55	4,01	100
Ano Entrada	EM	Não Gaussiana	2011,23	4,77	100
Tempo Escola	EI	Não Gaussiana	0,65	1,57	100
Tempo Escola	EFAI	Não Gaussiana	2,77	2,77	100
Tempo Escola	EFAF	Não Gaussiana	4,45	3,82	100
Tempo Escola	EM	Não Gaussiana	5,73	4,52	100
Idade	EI	Não Gaussiana	3,71	1,16	100
Idade	EFAI	Gaussiana	8,23	1,30	100
Idade	EFAF	Não Gaussiana	12,30	1,21	100
Idade	EM	Não Gaussiana	15,69	0,94	100
Tempo Atraso	EI	Não Gaussiana	-0,04	0,42	100
Tempo Atraso	EFAI	Gaussiana	-0,11	0,41	100
Tempo Atraso	EFAF	Não Gaussiana	-0,18	0,48	100
Tempo Atraso	EM	Não Gaussiana	-0,24	0,50	100
Desconto	EI	Gaussiana	0,13	3,07	65,46
Desconto	EFAI	Gaussiana	0,09	2,50	63,71
Desconto	EFAF	Gaussiana	0,10	2,43	63,14
Desconto	EM	Gaussiana	0,07	1,71	60,44
Qt Faltas	EI	Gaussiana	10,62	13,46	99,92
Qt Faltas	EFAI	Gaussiana	11,81	13,82	92,86
Qt Faltas	EFAF	Gaussiana	31,03	24,11	86,6
Qt Faltas	EM	Gaussiana	33,76	25,07	81,45
Média Nota	EI	Gaussiana	9,52	0,70	63,1
Média Nota	EFAI	Gaussiana	9,55	0,49	92,23
Média Nota	EFAF	Gaussiana	9,07	0,75	92,58
Média Nota	EM	Gaussiana	8,61	0,90	92,06
Menor Nota	EI	Gaussiana	8,22	1,85	63,1
Menor Nota	EFAI	Gaussiana	7,80	1,94	92,23
Menor Nota	EFAF	Gaussiana	6,87	2,05	92,58
Menor Nota	EM	Gaussiana	6,60	1,76	92,06
Maior Nota	EI	Gaussiana	9,98	0,23	63,1
Maior Nota	EFAI	Gaussiana	9,99	0,15	92,23
Maior Nota	EFAF	Gaussiana	9,97	0,24	92,58
Maior Nota	EM	Gaussiana	9,91	0,36	92,06
Nac	EI	Não Gaussiana	0,30	0,46	100
Nac	EFAI	Gaussiana	0,49	0,50	93,31

[1]Para a determinação da normalidade das colunas, foi adotado o teste de Shapiro-Wilk com o p-value $>= 0.05$.

Nac	EFAF	Gaussiana	0,41	0,49	92,79
Nac	EM	Gaussiana	0,25	0,43	92,12
Integral	EI	Não Gaussiana	0,06	0,23	100
Integral	EFAI	Gaussiana	0,08	0,27	93,31
Integral	EFAF	Gaussiana	0,03	0,18	92,79
Integral	EM	Gaussiana	0,00	0,05	92,12
Pastoral	EI	Não Gaussiana	0,02	0,15	100
Pastoral	EFAI	Gaussiana	0,10	0,30	93,31
Pastoral	EFAF	Gaussiana	0,19	0,39	92,79
Pastoral	EM	Gaussiana	0,14	0,34	92,12
Al Possui Pai	EI	Não Gaussiana	0,99	0,11	100
Al Possui Pai	EFAI	Não Gaussiana	0,98	0,13	100
Al Possui Pai	EFAF	Não Gaussiana	0,98	0,14	100
Al Possui Pai	EM	Não Gaussiana	0,98	0,16	100
Al Possui Mãe	EI	Não Gaussiana	1,00	0,03	100
Al Possui Mãe	EFAI	Não Gaussiana	1,00	0,05	100
Al Possui Mãe	EFAF	Não Gaussiana	1,00	0,06	100
Al Possui Mãe	EM	Não Gaussiana	0,99	0,07	100
Pai Faleceu	EI	Gaussiana	0,00	0,06	98,73
Pai Faleceu	EFAI	Gaussiana	0,00	0,07	98,26
Pai Faleceu	EFAF	Gaussiana	0,01	0,08	97,98
Pai Faleceu	EM	Gaussiana	0,01	0,07	97,53
Pai Tipo Religiao	EI	Gaussiana	8,63	2,34	59
Pai Tipo Religiao	EFAI	Gaussiana	8,60	2,27	61,43
Pai Tipo Religiao	EFAF	Gaussiana	8,51	2,01	67,36
Pai Tipo Religiao	EM	Gaussiana	8,49	1,97	67,89
Pai Tipo Escolaridade	EI	Gaussiana	9,11	1,55	62,22
Pai Tipo Escolaridade	EFAI	Gaussiana	9,00	1,55	60,67
Pai Tipo Escolaridade	EFAF	Gaussiana	8,86	1,62	61,4
Pai Tipo Escolaridade	EM	Gaussiana	8,72	1,67	58,4
Pai Tipo Profissao	EI	Gaussiana	203,74	132,54	75,25
Pai Tipo Profissao	EFAI	Gaussiana	203,79	139,39	77,78
Pai Tipo Profissao	EFAF	Gaussiana	205,53	141,43	80,47
Pai Tipo Profissao	EM	Gaussiana	204,13	137,51	82,71
Pai Pai Grau Parentesco	EI	Gaussiana	8,97	0,46	45,91
Pai Pai Grau Parentesco	EFAI	Gaussiana	8,99	0,28	47,47
Pai Pai Grau Parentesco	EFAF	Gaussiana	9,00	0,27	47,22
Pai Pai Grau Parentesco	EM	Gaussiana	8,99	0,28	49,14
Pai Possui Ctps	EI	Gaussiana	0,02	0,15	98,73
Pai Possui Ctps	EFAI	Gaussiana	0,02	0,12	98,26
Pai Possui Ctps	EFAF	Gaussiana	0,01	0,11	97,98
Pai Possui Ctps	EM	Gaussiana	0,01	0,11	97,53
Pai Possui Rg	EI	Gaussiana	0,45	0,50	98,73
Pai Possui Rg	EFAI	Gaussiana	0,51	0,50	98,26
Pai Possui Rg	EFAF	Gaussiana	0,53	0,50	97,98
Pai Possui Rg	EM	Gaussiana	0,53	0,50	97,53
Mãe Tipo Religiao	EI	Gaussiana	8,60	2,20	60,81
Mãe Tipo Religiao	EFAI	Gaussiana	8,59	2,13	63,41
Mãe Tipo Religiao	EFAF	Gaussiana	8,55	2,00	70,25
Mãe Tipo Religiao	EM	Gaussiana	8,53	1,96	71,83
Mãe Tipo Escolaridade	EI	Gaussiana	9,21	1,41	63,83
Mãe Tipo Escolaridade	EFAI	Gaussiana	9,14	1,34	62,44
Mãe Tipo Escolaridade	EFAF	Gaussiana	8,99	1,41	64,27
Mãe Tipo Escolaridade	EM	Gaussiana	8,84	1,53	61,97
Mãe Tipo Profissao	EI	Gaussiana	216,52	138,53	71,52
Mãe Tipo Profissao	EFAI	Gaussiana	212,79	143,58	77,13
Mãe Tipo Profissao	EFAF	Gaussiana	215,74	145,46	81,89
Mãe Tipo Profissao	EM	Gaussiana	215,71	140,71	85,16
Mãe Mãe Grau Parentesco	EI	Gaussiana	7,08	0,54	46,07
Mãe Mãe Grau Parentesco	EFAI	Gaussiana	7,02	0,34	48,02
Mãe Mãe Grau Parentesco	EFAF	Gaussiana	7,02	0,30	48,01
Mãe Mãe Grau Parentesco	EM	Gaussiana	7,02	0,32	50,39
Mãe Possui Ctps	EI	Gaussiana	0,06	0,24	99,88
Mãe Possui Ctps	EFAI	Gaussiana	0,04	0,21	99,74
Mãe Possui Ctps	EFAF	Gaussiana	0,04	0,19	99,58
Mãe Possui Ctps	EM	Gaussiana	0,03	0,18	99,44
Mãe Possui Rg	EI	Gaussiana	0,47	0,50	99,88
Mãe Possui Rg	EFAI	Gaussiana	0,53	0,50	99,74
Mãe Possui Rg	EFAF	Gaussiana	0,55	0,50	99,58

Mãe Possui Rg	EM	Gaussiana	0,57	0,50	99,44
Trimestre	EI	Gaussiana			0
Trimestre	EFAI	Gaussiana	1,49	1,12	93,89
Trimestre	EFAF	Gaussiana	1,49	1,12	99,42
Trimestre	EM	Gaussiana	1,48	1,11	99,75
Artes	EI	Gaussiana			0
Artes	EFAI	Gaussiana	9,43	0,78	82,61
Artes	EFAF	Gaussiana	8,51	1,22	91,76
Artes	EM	Gaussiana	7,99	1,42	70,57
Biologia	EI	Gaussiana			0
Biologia	EFAI	Gaussiana			0
Biologia	EFAF	Gaussiana			0
Biologia	EM	Gaussiana	7,42	1,32	84,6
Ciência	EI	Gaussiana			0
Ciência	EFAI	Gaussiana	8,90	0,94	82,67
Ciência	EFAF	Gaussiana	7,98	1,20	91,84
Ciência	EM	Gaussiana	7,79	1,09	6,13
Ed. Física	EI	Gaussiana			0
Ed. Física	EFAI	Gaussiana	9,50	0,70	82,6
Ed. Física	EFAF	Gaussiana	9,03	0,97	91,73
Ed. Física	EM	Gaussiana	8,74	1,23	88,96
Filosofia	EI	Gaussiana			0
Filosofia	EFAI	Gaussiana	9,22	0,73	1,57
Filosofia	EFAF	Gaussiana	8,24	1,23	20,3
Filosofia	EM	Gaussiana	7,79	1,40	84,34
Física	EI	Gaussiana			0
Física	EFAI	Gaussiana			0
Física	EFAF	Gaussiana			0
Física	EM	Gaussiana	7,23	1,47	84,6
Geografia	EI	Gaussiana			0
Geografia	EFAI	Gaussiana	8,86	0,96	82,66
Geografia	EFAF	Gaussiana	7,93	1,16	91,83
Geografia	EM	Gaussiana	7,65	1,28	89,86
História	EI	Gaussiana			0
História	EFAI	Gaussiana	8,87	0,97	82,66
História	EFAF	Gaussiana	7,84	1,22	91,82
História	EM	Gaussiana	7,63	1,35	89,85
Inglês	EI	Gaussiana			0
Inglês	EFAI	Gaussiana	9,16	0,93	82,64
Inglês	EFAF	Gaussiana	8,05	1,31	91,81
Inglês	EM	Gaussiana	7,83	1,51	81,37
Matemática	EI	Gaussiana			0
Matemática	EFAI	Gaussiana	8,68	0,97	82,67
Matemática	EFAF	Gaussiana	7,66	1,42	91,87
Matemática	EM	Gaussiana	7,08	1,58	89,9
Português	EI	Gaussiana			0
Português	EFAI	Gaussiana	8,65	0,92	82,67
Português	EFAF	Gaussiana	7,79	1,08	91,86
Português	EM	Gaussiana	7,39	1,22	90,22
Química	EI	Gaussiana			0
Química	EFAI	Gaussiana			0
Química	EFAF	Gaussiana			0
Química	EM	Gaussiana	7,44	1,47	84,61
Sociologia	EI	Gaussiana			0
Sociologia	EFAI	Gaussiana			0
Sociologia	EFAF	Gaussiana			0
Sociologia	EM	Gaussiana	7,74	1,39	83,25
IDHM	EI	Gaussiana	0,80	0,02	98,9
IDHM	EFAI	Gaussiana	0,80	0,02	92,44
IDHM	EFAF	Gaussiana	0,80	0,02	91,75
IDHM	EM	Gaussiana	0,80	0,03	91,03
IDHME	EI	Gaussiana	0,74	0,03	98,9
IDHME	EFAI	Gaussiana	0,74	0,03	92,44
IDHME	EFAF	Gaussiana	0,74	0,03	91,75
IDHME	EM	Gaussiana	0,74	0,03	91,03
IDHML	EI	Gaussiana	0,85	0,01	98,9
IDHML	EFAI	Gaussiana	0,85	0,01	92,44
IDHML	EFAF	Gaussiana	0,85	0,01	91,75
IDHML	EM	Gaussiana	0,86	0,01	91,03

IDHMR	EI	Gaussiana	0,81	0,04	98,9
IDHMR	EFAI	Gaussiana	0,82	0,04	92,44
IDHMR	EFAF	Gaussiana	0,82	0,04	91,75
IDHMR	EM	Gaussiana	0,82	0,04	91,03
Idx Escolaridade	EI	Gaussiana	0,68	0,05	98,9
Idx Escolaridade	EFAI	Gaussiana	0,68	0,05	92,44
Idx Escolaridade	EFAF	Gaussiana	0,68	0,06	91,75
Idx Escolaridade	EM	Gaussiana	0,68	0,06	91,03
Idx Freq Escolar	EI	Gaussiana	0,77	0,03	98,9
Idx Freq Escolar	EFAI	Gaussiana	0,77	0,03	92,44
Idx Freq Escolar	EFAF	Gaussiana	0,77	0,03	91,75
Idx Freq Escolar	EM	Gaussiana	0,77	0,03	91,03
Gini	EI	Gaussiana	0,54	0,06	98,9
Gini	EFAI	Gaussiana	0,55	0,06	92,44
Gini	EFAF	Gaussiana	0,55	0,06	91,75
Gini	EM	Gaussiana	0,55	0,06	91,03
Prop Pobreza Extrema	EI	Gaussiana	0,74	0,38	98,9
Prop Pobreza Extrema	EFAI	Gaussiana	0,75	0,41	92,44
Prop Pobreza Extrema	EFAF	Gaussiana	0,76	0,46	91,75
Prop Pobreza Extrema	EM	Gaussiana	0,80	0,61	91,03
Prop Pobreza Extrema Infantil	EI	Gaussiana	1,50	0,74	98,9
Prop Pobreza Extrema Infantil	EFAI	Gaussiana	1,50	0,75	92,44
Prop Pobreza Extrema Infantil	EFAF	Gaussiana	1,51	0,82	91,75
Prop Pobreza Extrema Infantil	EM	Gaussiana	1,59	1,00	91,03
Prop Pobreza	EI	Gaussiana	3,18	1,54	98,9
Prop Pobreza	EFAI	Gaussiana	3,18	1,58	92,44
Prop Pobreza	EFAF	Gaussiana	3,20	1,66	91,75
Prop Pobreza	EM	Gaussiana	3,33	1,91	91,03
Exp Vida	EI	Gaussiana	76,24	0,81	98,9
Exp Vida	EFAI	Gaussiana	76,27	0,75	92,44
Exp Vida	EFAF	Gaussiana	76,27	0,75	91,75
Exp Vida	EM	Gaussiana	76,32	0,76	91,03
Fecundidade	EI	Gaussiana	1,63	0,15	98,9
Fecundidade	EFAI	Gaussiana	1,62	0,15	92,44
Fecundidade	EFAF	Gaussiana	1,62	0,15	91,75
Fecundidade	EM	Gaussiana	1,64	0,16	91,03
Exp Anos Estudo	EI	Gaussiana	10,52	0,47	98,9
Exp Anos Estudo	EFAI	Gaussiana	10,50	0,47	92,44
Exp Anos Estudo	EFAF	Gaussiana	10,50	0,49	91,75
Exp Anos Estudo	EM	Gaussiana	10,49	0,50	91,03
PIB	EI	Gaussiana	133585170,52	217815601,52	98,9
PIB	EFAI	Gaussiana	179411823,29	252415929,21	92,44
PIB	EFAF	Gaussiana	184182166,90	253656255,57	91,75
PIB	EM	Gaussiana	189218090,94	249457086,54	91,03
PIBPC	EI	Gaussiana	49875,17	16464,83	98,9
PIBPC	EFAI	Gaussiana	51186,66	16329,74	92,44
PIBPC	EFAF	Gaussiana	51163,26	16348,12	91,75
PIBPC	EM	Gaussiana	52268,53	17040,03	91,03
Escola	EI	Não Gaussiana	0,49	0,50	100
Escola	EFAI	Não Gaussiana	0,49	0,50	100
Escola	EFAF	Não Gaussiana	0,50	0,50	100
Escola	EM	Não Gaussiana	0,52	0,50	100
Sexo Aluno	EI	Não Gaussiana	8,25	5,97	100
Sexo Aluno	EFAI	Não Gaussiana	5,77	4,87	100
Sexo Aluno	EFAF	Não Gaussiana	5,12	4,51	100
Sexo Aluno	EM	Não Gaussiana	5,95	4,96	100
Dominio Mãe	EI	Não Gaussiana	156,66	595,72	100
Dominio Mãe	EFAI	Não Gaussiana	148,17	578,10	100
Dominio Mãe	EFAF	Não Gaussiana	223,98	731,10	100
Dominio Mãe	EM	Não Gaussiana	320,99	881,39	100
Dominio Pai	EI	Não Gaussiana	672,96	1617,91	100
Dominio Pai	EFAI	Não Gaussiana	643,19	1626,63	100
Dominio Pai	EFAF	Não Gaussiana	844,11	1870,87	100
Dominio Pai	EM	Não Gaussiana	1178,24	2165,49	100
Servico	EI	Gaussiana	3,00	0,00	100
Servico	EFAI	Gaussiana	0,00	0,00	100
Servico	EFAF	Gaussiana	1,00	0,00	100
Servico	EM	Gaussiana	2,00	0,00	100
Serie	EI	Não Gaussiana	13,24	1,13	100

Serie	EFAI	Não Gaussiana	5,63	2,54	100
Serie	EFAF	Não Gaussiana	1,97	1,59	100
Serie	EM	Não Gaussiana	8,57	1,26	100
$N_{acumulado}$ Qt Faltas	EI	Não Gaussiana	10,68	13,50	100
$N_{acumulado}$ Qt Faltas	EFAI	Gaussiana	24,11	32,31	93,31
$N_{acumulado}$ Qt Faltas	EFAF	Gaussiana	62,55	61,18	92,79
$N_{acumulado}$ Qt Faltas	EM	Gaussiana	65,87	66,10	92,12
$N_{acumulado}$ Artes	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Artes	EFAI	Gaussiana	17,90	9,89	93,31
$N_{acumulado}$ Artes	EFAF	Gaussiana	18,53	7,71	92,79
$N_{acumulado}$ Artes	EM	Gaussiana	13,67	9,71	92,12
$N_{acumulado}$ Biologia	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Biologia	EFAI	Gaussiana	0,00	0,00	93,31
$N_{acumulado}$ Biologia	EFAF	Gaussiana	0,00	0,00	92,79
$N_{acumulado}$ Biologia	EM	Gaussiana	14,69	7,58	92,12
$N_{acumulado}$ Ciência	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Ciência	EFAI	Gaussiana	16,97	9,42	93,31
$N_{acumulado}$ Ciência	EFAF	Gaussiana	17,36	7,37	92,79
$N_{acumulado}$ Ciência	EM	Gaussiana	1,50	5,71	92,12
$N_{acumulado}$ Filosofia	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Filosofia	EFAI	Gaussiana	0,34	2,77	93,31
$N_{acumulado}$ Filosofia	EFAF	Gaussiana	3,88	8,07	92,79
$N_{acumulado}$ Filosofia	EM	Gaussiana	15,50	7,97	92,12
$N_{acumulado}$ Física	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Física	EFAI	Gaussiana	0,00	0,00	93,31
$N_{acumulado}$ Física	EFAF	Gaussiana	0,00	0,00	92,79
$N_{acumulado}$ Física	EM	Gaussiana	14,32	7,53	92,12
$N_{acumulado}$ Geografia	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Geografia	EFAI	Gaussiana	16,87	9,38	93,31
$N_{acumulado}$ Geografia	EFAF	Gaussiana	17,25	7,25	92,79
$N_{acumulado}$ Geografia	EM	Gaussiana	16,48	7,02	92,12
$N_{acumulado}$ História	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ História	EFAI	Gaussiana	16,91	9,40	93,31
$N_{acumulado}$ História	EFAF	Gaussiana	17,11	7,20	92,79
$N_{acumulado}$ História	EM	Gaussiana	16,44	7,05	92,12
$N_{acumulado}$ Inglês	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Inglês	EFAI	Gaussiana	17,43	9,66	93,31
$N_{acumulado}$ Inglês	EFAF	Gaussiana	17,54	7,48	92,79
$N_{acumulado}$ Inglês	EM	Gaussiana	15,28	8,41	92,12
$N_{acumulado}$ Português	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Português	EFAI	Gaussiana	16,50	9,15	93,31
$N_{acumulado}$ Português	EFAF	Gaussiana	16,94	7,10	92,79
$N_{acumulado}$ Português	EM	Gaussiana	16,04	6,78	92,12
$N_{acumulado}$ Química	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Química	EFAI	Gaussiana	0,00	0,00	93,31
$N_{acumulado}$ Química	EFAF	Gaussiana	0,00	0,00	92,79
$N_{acumulado}$ Química	EM	Gaussiana	14,77	7,70	92,12
$N_{acumulado}$ Sociologia	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Sociologia	EFAI	Gaussiana	0,00	0,00	93,31
$N_{acumulado}$ Sociologia	EFAF	Gaussiana	0,00	0,00	92,79
$N_{acumulado}$ Sociologia	EM	Gaussiana	15,04	8,03	92,12
$N_{acumulado}$ Ed. Física	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Ed. Física	EFAI	Gaussiana	18,05	9,93	93,31
$N_{acumulado}$ Ed. Física	EFAF	Gaussiana	19,60	8,01	92,79
$N_{acumulado}$ Ed. Física	EM	Gaussiana	18,68	7,95	92,12
$N_{acumulado}$ Matemática	EI	Gaussiana	0,00	0,00	100
$N_{acumulado}$ Matemática	EFAI	Gaussiana	16,56	9,20	93,31
$N_{acumulado}$ Matemática	EFAF	Gaussiana	16,69	7,28	92,79
$N_{acumulado}$ Matemática	EM	Gaussiana	15,27	6,95	92,12
$\Delta N_{trimestral}$ Qt Faltas	EI	Não Gaussiana	-0,01	1,01	100
$\Delta N_{trimestral}$ Qt Faltas	EFAI	Gaussiana	0,00	0,38	93,31
$\Delta N_{trimestral}$ Qt Faltas	EFAF	Gaussiana	0,01	1,05	92,79
$\Delta N_{trimestral}$ Qt Faltas	EM	Gaussiana	0,01	0,94	92,12
$\Delta N_{trimestral}$ Artes	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Artes	EFAI	Gaussiana	0,03	0,57	93,31
$\Delta N_{trimestral}$ Artes	EFAF	Gaussiana	-0,03	1,00	92,79
$\Delta N_{trimestral}$ Artes	EM	Gaussiana	-0,03	0,99	92,12
$\Delta N_{trimestral}$ Biologia	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Biologia	EFAI	Gaussiana	0,00	0,00	93,31

$\Delta N_{trimestral}$ Biologia	EFAF	Gaussiana	0,00	0,00	92,79
$\Delta N_{trimestral}$ Biologia	EM	Gaussiana	-0,02	0,92	92,12
$\Delta N_{trimestral}$ Ciência	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Ciência	EFAI	Gaussiana	0,00	0,66	93,31
$\Delta N_{trimestral}$ Ciência	EFAF	Gaussiana	0,01	0,80	92,79
$\Delta N_{trimestral}$ Ciência	EM	Gaussiana	0,00	0,24	92,12
$\Delta N_{trimestral}$ Ed. Física	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Ed. Física	EFAI	Gaussiana	0,02	0,48	93,31
$\Delta N_{trimestral}$ Ed. Física	EFAF	Gaussiana	0,00	0,75	92,79
$\Delta N_{trimestral}$ Ed. Física	EM	Gaussiana	-0,05	0,94	92,12
$\Delta N_{trimestral}$ Filosofia	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Filosofia	EFAI	Gaussiana	0,00	0,09	93,31
$\Delta N_{trimestral}$ Filosofia	EFAF	Gaussiana	-0,01	0,50	92,79
$\Delta N_{trimestral}$ Filosofia	EM	Gaussiana	-0,02	1,09	92,12
$\Delta N_{trimestral}$ Física	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Física	EFAI	Gaussiana	0,00	0,00	93,31
$\Delta N_{trimestral}$ Física	EFAF	Gaussiana	0,00	0,00	92,79
$\Delta N_{trimestral}$ Física	EM	Gaussiana	-0,03	1,00	92,12
$\Delta N_{trimestral}$ Geografia	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Geografia	EFAI	Gaussiana	0,01	0,70	93,31
$\Delta N_{trimestral}$ Geografia	EFAF	Gaussiana	0,02	0,88	92,79
$\Delta N_{trimestral}$ Geografia	EM	Gaussiana	-0,03	0,99	92,12
$\Delta N_{trimestral}$ História	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ História	EFAI	Gaussiana	-0,01	0,68	93,31
$\Delta N_{trimestral}$ História	EFAF	Gaussiana	-0,04	0,90	92,79
$\Delta N_{trimestral}$ História	EM	Gaussiana	-0,04	1,02	92,12
$\Delta N_{trimestral}$ Inglês	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Inglês	EFAI	Gaussiana	0,01	0,67	93,31
$\Delta N_{trimestral}$ Inglês	EFAF	Gaussiana	0,00	0,87	92,79
$\Delta N_{trimestral}$ Inglês	EM	Gaussiana	0,00	1,12	92,12
$\Delta N_{trimestral}$ Matemática	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Matemática	EFAI	Gaussiana	-0,03	0,60	93,31
$\Delta N_{trimestral}$ Matemática	EFAF	Gaussiana	0,00	0,87	92,79
$\Delta N_{trimestral}$ Matemática	EM	Gaussiana	-0,01	1,01	92,12
$\Delta N_{trimestral}$ Português	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Português	EFAI	Gaussiana	-0,02	0,56	93,31
$\Delta N_{trimestral}$ Português	EFAF	Gaussiana	0,01	0,66	92,79
$\Delta N_{trimestral}$ Português	EM	Gaussiana	-0,05	0,80	92,12
$\Delta N_{trimestral}$ Química	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Química	EFAI	Gaussiana	0,00	0,00	93,31
$\Delta N_{trimestral}$ Química	EFAF	Gaussiana	0,00	0,00	92,79
$\Delta N_{trimestral}$ Química	EM	Gaussiana	-0,05	0,99	92,12
$\Delta N_{trimestral}$ Sociologia	EI	Gaussiana	0,00	0,00	100
$\Delta N_{trimestral}$ Sociologia	EFAI	Gaussiana	0,00	0,00	93,31
$\Delta N_{trimestral}$ Sociologia	EFAF	Gaussiana	0,00	0,00	92,79
$\Delta N_{trimestral}$ Sociologia	EM	Gaussiana	0,02	1,09	92,12
ΔN_{anual} Anuidade	EI	Não Gaussiana	1734,90	4056,87	100
ΔN_{anual} Anuidade	EFAI	Gaussiana	1063,44	3405,26	93,31
ΔN_{anual} Anuidade	EFAF	Gaussiana	1045,96	3619,07	92,79
ΔN_{anual} Anuidade	EM	Gaussiana	1175,31	4073,62	92,12
ΔN_{anual} Desconto	EI	Não Gaussiana	0,01	1,83	100
ΔN_{anual} Desconto	EFAI	Gaussiana	0,01	1,18	93,31
ΔN_{anual} Desconto	EFAF	Gaussiana	0,01	1,11	92,79
ΔN_{anual} Desconto	EM	Gaussiana	0,01	0,82	92,12
ΔN_{anual} Qt Faltas	EI	Não Gaussiana	0,82	8,46	100
ΔN_{anual} Qt Faltas	EFAI	Gaussiana	0,64	7,32	93,31
ΔN_{anual} Qt Faltas	EFAF	Gaussiana	2,11	11,36	92,79
ΔN_{anual} Qt Faltas	EM	Gaussiana	1,37	11,14	92,12
ΔN_{anual} Artes	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Artes	EFAI	Gaussiana	-0,04	0,57	93,31
ΔN_{anual} Artes	EFAF	Gaussiana	-0,17	1,04	92,79
ΔN_{anual} Artes	EM	Gaussiana	-0,04	0,98	92,12
ΔN_{anual} Biologia	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Biologia	EFAI	Gaussiana	0,00	0,00	93,31
ΔN_{anual} Biologia	EFAF	Gaussiana	0,00	0,00	92,79
ΔN_{anual} Biologia	EM	Gaussiana	-0,06	0,84	92,12
ΔN_{anual} Ciência	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Ciência	EFAI	Gaussiana	-0,11	0,66	93,31
ΔN_{anual} Ciência	EFAF	Gaussiana	-0,17	0,91	92,79

ΔN_{anual} Ciência	EM	Gaussiana	0,00	0,25	92,12
ΔN_{anual} Ed. Física	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Ed. Física	EFAI	Gaussiana	-0,01	0,46	93,31
ΔN_{anual} Ed. Física	EFAF	Gaussiana	-0,08	0,80	92,79
ΔN_{anual} Ed. Física	EM	Gaussiana	0,02	0,98	92,12
ΔN_{anual} Filosofia	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Filosofia	EFAI	Gaussiana	0,00	0,06	93,31
ΔN_{anual} Filosofia	EFAF	Gaussiana	0,00	0,40	92,79
ΔN_{anual} Filosofia	EM	Gaussiana	-0,04	0,98	92,12
ΔN_{anual} Física	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Física	EFAI	Gaussiana	0,00	0,00	93,31
ΔN_{anual} Física	EFAF	Gaussiana	0,00	0,00	92,79
ΔN_{anual} Física	EM	Gaussiana	-0,04	0,91	92,12
ΔN_{anual} Geografia	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Geografia	EFAI	Gaussiana	-0,12	0,68	93,31
ΔN_{anual} Geografia	EFAF	Gaussiana	-0,14	0,92	92,79
ΔN_{anual} Geografia	EM	Gaussiana	-0,04	1,02	92,12
ΔN_{anual} História	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} História	EFAI	Gaussiana	-0,13	0,68	93,31
ΔN_{anual} História	EFAF	Gaussiana	-0,15	0,97	92,79
ΔN_{anual} História	EM	Gaussiana	-0,04	1,03	92,12
ΔN_{anual} Inglês	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Inglês	EFAI	Gaussiana	-0,09	0,65	93,31
ΔN_{anual} Inglês	EFAF	Gaussiana	-0,17	0,95	92,79
ΔN_{anual} Inglês	EM	Gaussiana	0,02	1,11	92,12
ΔN_{anual} Matemática	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Matemática	EFAI	Gaussiana	-0,11	0,60	93,31
ΔN_{anual} Matemática	EFAF	Gaussiana	-0,17	0,94	92,79
ΔN_{anual} Matemática	EM	Gaussiana	-0,09	1,08	92,12
ΔN_{anual} Português	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Português	EFAI	Gaussiana	-0,07	0,57	93,31
ΔN_{anual} Português	EFAF	Gaussiana	-0,12	0,75	92,79
ΔN_{anual} Português	EM	Gaussiana	-0,07	0,85	92,12
ΔN_{anual} Química	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Química	EFAI	Gaussiana	0,00	0,00	93,31
ΔN_{anual} Química	EFAF	Gaussiana	0,00	0,00	92,79
ΔN_{anual} Química	EM	Gaussiana	-0,06	0,89	92,12
ΔN_{anual} Sociologia	EI	Gaussiana	0,00	0,00	100
ΔN_{anual} Sociologia	EFAI	Gaussiana	0,00	0,00	93,31
ΔN_{anual} Sociologia	EFAF	Gaussiana	0,00	0,00	92,79
ΔN_{anual} Sociologia	EM	Gaussiana	0,00	0,92	92,12

Apêndice C

Distribuição das colunas por trimestre

Tabela C.1: Distribuição das colunas por trimestre

Nome da Coluna	Trimestre	Distribuição[1]	Média	Desvio Padrão	Não-nulos (%)
Anuidade	1	Não Gaussiana	13801,61	8813,74	100
Anuidade	2	Não Gaussiana	13817,83	8838,59	100
Anuidade	3	Não Gaussiana	13876,09	8764,71	100
Ano Entrada	1	Não Gaussiana	2012,83	4,05	100
Ano Entrada	2	Não Gaussiana	2012,82	4,05	100
Ano Entrada	3	Não Gaussiana	2012,82	4,06	100
Tempo Escola	1	Não Gaussiana	4,16	3,84	100
Tempo Escola	2	Não Gaussiana	4,16	3,84	100
Tempo Escola	3	Não Gaussiana	4,18	3,85	100
Idade	1	Não Gaussiana	11,66	3,10	100
Idade	2	Não Gaussiana	11,65	3,09	100
Idade	3	Não Gaussiana	11,66	3,08	100
Tempo Atraso	1	Não Gaussiana	-0,17	0,46	100
Tempo Atraso	2	Não Gaussiana	-0,17	0,46	100
Tempo Atraso	3	Não Gaussiana	-0,17	0,46	100
Desconto	1	Gaussiana	0,10	2,43	67,26
Desconto	2	Gaussiana	0,09	2,36	67,2
Desconto	3	Gaussiana	0,09	2,33	66,87
Qt Faltas	1	Gaussiana	24,87	23,45	93,95
Qt Faltas	2	Gaussiana	25,01	23,50	93,9
Qt Faltas	3	Gaussiana	25,07	23,48	93,91
Média Nota	1	Gaussiana	9,09	0,86	99,98
Média Nota	2	Gaussiana	9,11	0,83	100
Média Nota	3	Gaussiana	9,13	0,79	100
Menor Nota	1	Gaussiana	7,09	2,04	99,98
Menor Nota	2	Gaussiana	7,10	2,03	100
Menor Nota	3	Gaussiana	7,13	2,00	100
Maior Nota	1	Gaussiana	9,95	0,31	99,98
Maior Nota	2	Gaussiana	9,96	0,25	100
Maior Nota	3	Gaussiana	9,96	0,22	100
Nac	1	Não Gaussiana	0,40	0,49	100
Nac	2	Não Gaussiana	0,40	0,49	100
Nac	3	Não Gaussiana	0,40	0,49	100
Integral	1	Não Gaussiana	0,04	0,20	100
Integral	2	Não Gaussiana	0,04	0,20	100
Integral	3	Não Gaussiana	0,04	0,20	100
Pastoral	1	Não Gaussiana	0,15	0,36	100
Pastoral	2	Não Gaussiana	0,15	0,36	100
Pastoral	3	Não Gaussiana	0,15	0,36	100
Al Possui Pai	1	Não Gaussiana	0,98	0,14	100
Al Possui Pai	2	Não Gaussiana	0,98	0,14	100
Al Possui Pai	3	Não Gaussiana	0,98	0,14	100
Al Possui Mãe	1	Não Gaussiana	1,00	0,06	100
Al Possui Mãe	2	Não Gaussiana	1,00	0,06	100
Al Possui Mãe	3	Não Gaussiana	1,00	0,06	100
Pai Faleceu	1	Gaussiana	0,01	0,07	97,94
Pai Faleceu	2	Gaussiana	0,01	0,07	97,96
Pai Faleceu	3	Gaussiana	0,01	0,07	97,99
Pai Tipo Religiao	1	Gaussiana	8,54	2,10	65,4
Pai Tipo Religiao	2	Gaussiana	8,54	2,10	65,55
Pai Tipo Religiao	3	Gaussiana	8,53	2,10	65,53

[1]Para a determinação da normalidade das colunas, foi adotado o teste de Shapiro-Wilk com o p-value $>= 0.05$.

Pai Tipo Escolaridade	1	Gaussiana	8,88	1,61	60,5
Pai Tipo Escolaridade	2	Gaussiana	8,88	1,61	60,76
Pai Tipo Escolaridade	3	Gaussiana	8,87	1,60	60,62
Pai Tipo Profissao	1	Gaussiana	204,53	139,75	80,15
Pai Tipo Profissao	2	Gaussiana	204,68	139,86	80,22
Pai Tipo Profissao	3	Gaussiana	205,15	139,92	80,16
Pai Pai Grau Parentesco	1	Gaussiana	8,99	0,25	47,76
Pai Pai Grau Parentesco	2	Gaussiana	8,99	0,25	47,54
Pai Pai Grau Parentesco	3	Gaussiana	8,99	0,25	47
Pai Possui Ctps	1	Gaussiana	0,01	0,11	97,94
Pai Possui Ctps	2	Gaussiana	0,01	0,11	97,96
Pai Possui Ctps	3	Gaussiana	0,01	0,11	97,99
Pai Possui Rg	1	Gaussiana	0,52	0,50	97,94
Pai Possui Rg	2	Gaussiana	0,52	0,50	97,96
Pai Possui Rg	3	Gaussiana	0,52	0,50	97,99
Mãe Tipo Religiao	1	Gaussiana	8,56	2,04	68,23
Mãe Tipo Religiao	2	Gaussiana	8,56	2,04	68,35
Mãe Tipo Religiao	3	Gaussiana	8,56	2,03	68,33
Mãe Tipo Escolaridade	1	Gaussiana	9,01	1,42	63,14
Mãe Tipo Escolaridade	2	Gaussiana	9,01	1,42	63,39
Mãe Tipo Escolaridade	3	Gaussiana	9,01	1,42	63,24
Mãe Tipo Profissao	1	Gaussiana	214,57	143,67	81,12
Mãe Tipo Profissao	2	Gaussiana	214,69	143,65	81,17
Mãe Tipo Profissao	3	Gaussiana	215,15	143,52	81,08
Mãe Mãe Grau Parentesco	1	Gaussiana	7,02	0,28	48,59
Mãe Mãe Grau Parentesco	2	Gaussiana	7,02	0,28	48,36
Mãe Mãe Grau Parentesco	3	Gaussiana	7,02	0,28	47,8
Mãe Possui Ctps	1	Gaussiana	0,04	0,19	99,6
Mãe Possui Ctps	2	Gaussiana	0,04	0,19	99,6
Mãe Possui Ctps	3	Gaussiana	0,04	0,19	99,61
Mãe Possui Rg	1	Gaussiana	0,55	0,50	99,6
Mãe Possui Rg	2	Gaussiana	0,55	0,50	99,6
Mãe Possui Rg	3	Gaussiana	0,55	0,50	99,61
Trimestre	1	Gaussiana	1,00	0,00	100
Trimestre	2	Gaussiana	2,00	0,00	100
Trimestre	3	Gaussiana	3,00	0,00	100
Artes	1	Gaussiana	8,74	1,17	93,34
Artes	2	Gaussiana	8,71	1,31	93,17
Artes	3	Gaussiana	8,76	1,32	93,62
Biologia	1	Gaussiana	7,41	1,20	24,36
Biologia	2	Gaussiana	7,31	1,41	24,15
Biologia	3	Gaussiana	7,46	1,36	24,11
Ciência	1	Gaussiana	8,39	1,16	74,89
Ciência	2	Gaussiana	8,36	1,23	75,16
Ciência	3	Gaussiana	8,43	1,17	75,41
Ed. Física	1	Gaussiana	9,10	0,93	98,75
Ed. Física	2	Gaussiana	9,12	1,07	98,54
Ed. Física	3	Gaussiana	9,14	1,05	98,99
Filosofia	1	Gaussiana	7,92	1,25	33,26
Filosofia	2	Gaussiana	7,88	1,44	33,04
Filosofia	3	Gaussiana	7,93	1,45	33,15
Física	1	Gaussiana	7,19	1,40	24,36
Física	2	Gaussiana	7,18	1,55	24,17
Física	3	Gaussiana	7,24	1,49	24,09
Geografia	1	Gaussiana	8,18	1,19	99,01
Geografia	2	Gaussiana	8,12	1,29	99,07
Geografia	3	Gaussiana	8,23	1,26	99,28
História	1	Gaussiana	8,19	1,22	99,02
História	2	Gaussiana	8,10	1,35	99,05
História	3	Gaussiana	8,15	1,31	99,29
Inglês	1	Gaussiana	8,40	1,30	96,83
Inglês	2	Gaussiana	8,34	1,47	96,74
Inglês	3	Gaussiana	8,44	1,37	96,81
Matemática	1	Gaussiana	7,90	1,43	99,03
Matemática	2	Gaussiana	7,78	1,56	99,16
Matemática	3	Gaussiana	7,90	1,46	99,29
Português	1	Gaussiana	8,01	1,12	99,19
Português	2	Gaussiana	7,93	1,25	99,17
Português	3	Gaussiana	8,00	1,21	99,31

Química	1	Gaussiana	7,48	1,36	24,36
Química	2	Gaussiana	7,33	1,54	24,18
Química	3	Gaussiana	7,44	1,55	24,1
Sociologia	1	Gaussiana	7,65	1,27	24,01
Sociologia	2	Gaussiana	7,67	1,45	23,71
Sociologia	3	Gaussiana	7,84	1,45	23,74
IDHM	1	Gaussiana	0,80	0,02	98,91
IDHM	2	Gaussiana	0,80	0,02	98,92
IDHM	3	Gaussiana	0,80	0,02	98,92
IDHME	1	Gaussiana	0,74	0,03	98,91
IDHME	2	Gaussiana	0,74	0,03	98,92
IDHME	3	Gaussiana	0,74	0,03	98,92
IDHML	1	Gaussiana	0,85	0,01	98,91
IDHML	2	Gaussiana	0,85	0,01	98,92
IDHML	3	Gaussiana	0,85	0,01	98,92
IDHMR	1	Gaussiana	0,82	0,04	98,91
IDHMR	2	Gaussiana	0,82	0,04	98,92
IDHMR	3	Gaussiana	0,82	0,04	98,92
Idx Escolaridade	1	Gaussiana	0,68	0,06	98,91
Idx Escolaridade	2	Gaussiana	0,68	0,06	98,92
Idx Escolaridade	3	Gaussiana	0,68	0,06	98,92
Idx Freq Escolar	1	Gaussiana	0,77	0,03	98,91
Idx Freq Escolar	2	Gaussiana	0,77	0,03	98,92
Idx Freq Escolar	3	Gaussiana	0,77	0,03	98,92
Gini	1	Gaussiana	0,55	0,06	98,91
Gini	2	Gaussiana	0,55	0,06	98,92
Gini	3	Gaussiana	0,55	0,06	98,92
Prop Pobreza Extrema	1	Gaussiana	0,77	0,50	98,91
Prop Pobreza Extrema	2	Gaussiana	0,77	0,50	98,92
Prop Pobreza Extrema	3	Gaussiana	0,77	0,49	98,92
Prop Pobreza Extrema Infantil	1	Gaussiana	1,53	0,85	98,91
Prop Pobreza Extrema Infantil	2	Gaussiana	1,53	0,86	98,92
Prop Pobreza Extrema Infantil	3	Gaussiana	1,53	0,85	98,92
Prop Pobreza	1	Gaussiana	3,24	1,71	98,91
Prop Pobreza	2	Gaussiana	3,25	1,71	98,92
Prop Pobreza	3	Gaussiana	3,23	1,71	98,92
Exp Vida	1	Gaussiana	76,28	0,75	98,91
Exp Vida	2	Gaussiana	76,28	0,75	98,92
Exp Vida	3	Gaussiana	76,28	0,76	98,92
Fecundidade	1	Gaussiana	1,63	0,15	98,91
Fecundidade	2	Gaussiana	1,63	0,15	98,92
Fecundidade	3	Gaussiana	1,63	0,15	98,92
Exp Anos Estudo	1	Gaussiana	10,49	0,49	98,91
Exp Anos Estudo	2	Gaussiana	10,49	0,49	98,92
Exp Anos Estudo	3	Gaussiana	10,50	0,49	98,92
PIB	1	Gaussiana	185549193,26	253536336,62	98,91
PIB	2	Gaussiana	185896418,36	253752337,40	98,92
PIB	3	Gaussiana	181658004,52	250584222,23	98,92
PIBPC	1	Gaussiana	51534,23	16582,55	98,91
PIBPC	2	Gaussiana	51567,54	16622,48	98,92
PIBPC	3	Gaussiana	51466,25	16646,08	98,92
Escola	1	Não Gaussiana	0,50	0,50	100
Escola	2	Não Gaussiana	0,50	0,50	100
Escola	3	Não Gaussiana	0,50	0,50	100
Sexo Aluno	1	Não Gaussiana	5,48	4,68	100
Sexo Aluno	2	Não Gaussiana	5,49	4,69	100
Sexo Aluno	3	Não Gaussiana	5,52	4,68	100
Dominio Mãe	1	Não Gaussiana	220,43	725,67	100
Dominio Mãe	2	Não Gaussiana	218,39	721,38	100
Dominio Mãe	3	Não Gaussiana	216,48	718,03	100
Dominio Pai	1	Não Gaussiana	855,09	1880,23	100
Dominio Pai	2	Não Gaussiana	847,24	1871,46	100
Dominio Pai	3	Não Gaussiana	840,09	1864,06	100
Servico	1	Não Gaussiana	0,87	0,78	100
Servico	2	Não Gaussiana	0,87	0,78	100
Servico	3	Não Gaussiana	0,87	0,78	100
Serie	1	Não Gaussiana	4,88	3,14	100
Serie	2	Não Gaussiana	4,85	3,13	100
Serie	3	Não Gaussiana	4,82	3,13	100

$N_{acumulado}$ Qt Faltas	1	Não Gaussiana	23,75	24,13	100
$N_{acumulado}$ Qt Faltas	2	Não Gaussiana	47,38	47,76	100
$N_{acumulado}$ Qt Faltas	3	Não Gaussiana	71,14	71,45	100
$N_{acumulado}$ Artes	1	Não Gaussiana	8,20	2,44	100
$N_{acumulado}$ Artes	2	Não Gaussiana	16,40	4,68	100
$N_{acumulado}$ Artes	3	Não Gaussiana	24,70	6,89	100
$N_{acumulado}$ Biologia	1	Não Gaussiana	1,85	3,27	100
$N_{acumulado}$ Biologia	2	Não Gaussiana	3,62	6,45	100
$N_{acumulado}$ Biologia	3	Não Gaussiana	5,45	9,72	100
$N_{acumulado}$ Ciência	1	Não Gaussiana	6,31	3,78	100
$N_{acumulado}$ Ciência	2	Não Gaussiana	12,65	7,47	100
$N_{acumulado}$ Ciência	3	Não Gaussiana	19,08	11,16	100
$N_{acumulado}$ Filosofia	1	Não Gaussiana	2,68	3,82	100
$N_{acumulado}$ Filosofia	2	Não Gaussiana	5,30	7,57	100
$N_{acumulado}$ Filosofia	3	Não Gaussiana	7,98	11,39	100
$N_{acumulado}$ Física	1	Não Gaussiana	1,79	3,20	100
$N_{acumulado}$ Física	2	Não Gaussiana	3,54	6,33	100
$N_{acumulado}$ Física	3	Não Gaussiana	5,31	9,52	100
$N_{acumulado}$ Geografia	1	Não Gaussiana	8,17	1,31	100
$N_{acumulado}$ Geografia	2	Não Gaussiana	16,28	2,38	100
$N_{acumulado}$ Geografia	3	Não Gaussiana	24,54	3,30	100
$N_{acumulado}$ História	1	Não Gaussiana	8,17	1,34	100
$N_{acumulado}$ História	2	Não Gaussiana	16,26	2,48	100
$N_{acumulado}$ História	3	Não Gaussiana	24,45	3,45	100
$N_{acumulado}$ Inglês	1	Não Gaussiana	8,20	1,86	100
$N_{acumulado}$ Inglês	2	Não Gaussiana	16,32	3,60	100
$N_{acumulado}$ Inglês	3	Não Gaussiana	24,60	5,23	100
$N_{acumulado}$ Português	1	Não Gaussiana	8,03	1,32	100
$N_{acumulado}$ Português	2	Não Gaussiana	15,93	2,33	100
$N_{acumulado}$ Português	3	Não Gaussiana	23,98	3,28	100
$N_{acumulado}$ Química	1	Não Gaussiana	1,86	3,31	100
$N_{acumulado}$ Química	2	Não Gaussiana	3,65	6,52	100
$N_{acumulado}$ Química	3	Não Gaussiana	5,47	9,79	100
$N_{acumulado}$ Sociologia	1	Não Gaussiana	1,88	3,36	100
$N_{acumulado}$ Sociologia	2	Não Gaussiana	3,71	6,66	100
$N_{acumulado}$ Sociologia	3	Não Gaussiana	5,60	10,07	100
$N_{acumulado}$ Ed. Física	1	Não Gaussiana	9,07	1,21	100
$N_{acumulado}$ Ed. Física	2	Não Gaussiana	18,11	2,21	100
$N_{acumulado}$ Ed. Física	3	Não Gaussiana	27,27	3,05	100
$N_{acumulado}$ Matemática	1	Não Gaussiana	7,88	1,52	100
$N_{acumulado}$ Matemática	2	Não Gaussiana	15,67	2,91	100
$N_{acumulado}$ Matemática	3	Não Gaussiana	23,62	4,07	100
$\Delta N_{trimestral}$ Qt Faltas	1	Não Gaussiana	0,02	1,26	100
$\Delta N_{trimestral}$ Qt Faltas	2	Não Gaussiana	0,00	0,79	100
$\Delta N_{trimestral}$ Qt Faltas	3	Não Gaussiana	0,00	0,41	100
$\Delta N_{trimestral}$ Artes	1	Não Gaussiana	0,00	0,05	100
$\Delta N_{trimestral}$ Artes	2	Não Gaussiana	-0,04	1,05	100
$\Delta N_{trimestral}$ Artes	3	Não Gaussiana	0,01	1,03	100
$\Delta N_{trimestral}$ Biologia	1	Não Gaussiana	0,00	0,03	100
$\Delta N_{trimestral}$ Biologia	2	Não Gaussiana	-0,03	0,58	100
$\Delta N_{trimestral}$ Biologia	3	Não Gaussiana	0,01	0,55	100
$\Delta N_{trimestral}$ Ciência	1	Não Gaussiana	0,00	0,06	100
$\Delta N_{trimestral}$ Ciência	2	Não Gaussiana	-0,03	0,81	100
$\Delta N_{trimestral}$ Ciência	3	Não Gaussiana	0,03	0,76	100
$\Delta N_{trimestral}$ Ed. Física	1	Não Gaussiana	0,00	0,06	100
$\Delta N_{trimestral}$ Ed. Física	2	Não Gaussiana	0,00	0,94	100
$\Delta N_{trimestral}$ Ed. Física	3	Não Gaussiana	-0,01	0,83	100
$\Delta N_{trimestral}$ Filosofia	1	Não Gaussiana	0,00	0,03	100
$\Delta N_{trimestral}$ Filosofia	2	Não Gaussiana	-0,02	0,76	100
$\Delta N_{trimestral}$ Filosofia	3	Não Gaussiana	-0,01	0,76	100
$\Delta N_{trimestral}$ Física	1	Não Gaussiana	0,00	0,04	100
$\Delta N_{trimestral}$ Física	2	Não Gaussiana	-0,01	0,64	100
$\Delta N_{trimestral}$ Física	3	Não Gaussiana	-0,01	0,59	100
$\Delta N_{trimestral}$ Geografia	1	Não Gaussiana	0,00	0,07	100
$\Delta N_{trimestral}$ Geografia	2	Não Gaussiana	-0,08	1,06	100
$\Delta N_{trimestral}$ Geografia	3	Não Gaussiana	0,06	1,00	100
$\Delta N_{trimestral}$ História	1	Não Gaussiana	0,00	0,06	100
$\Delta N_{trimestral}$ História	2	Não Gaussiana	-0,10	1,06	100
$\Delta N_{trimestral}$ História	3	Não Gaussiana	0,01	1,01	100

$\Delta N_{trimestral}$ Inglês	1	Não Gaussiana	0,00	0,07	100
$\Delta N_{trimestral}$ Inglês	2	Não Gaussiana	-0,07	1,13	100
$\Delta N_{trimestral}$ Inglês	3	Não Gaussiana	0,04	1,00	100
$\Delta N_{trimestral}$ Matemática	1	Não Gaussiana	0,00	0,06	100
$\Delta N_{trimestral}$ Matemática	2	Não Gaussiana	-0,13	1,02	100
$\Delta N_{trimestral}$ Matemática	3	Não Gaussiana	0,06	0,97	100
$\Delta N_{trimestral}$ Português	1	Não Gaussiana	0,00	0,07	100
$\Delta N_{trimestral}$ Português	2	Não Gaussiana	-0,09	0,87	100
$\Delta N_{trimestral}$ Português	3	Não Gaussiana	0,02	0,75	100
$\Delta N_{trimestral}$ Química	1	Não Gaussiana	0,00	0,03	100
$\Delta N_{trimestral}$ Química	2	Não Gaussiana	-0,04	0,63	100
$\Delta N_{trimestral}$ Química	3	Não Gaussiana	0,00	0,60	100
$\Delta N_{trimestral}$ Sociologia	1	Não Gaussiana	0,00	0,04	100
$\Delta N_{trimestral}$ Sociologia	2	Não Gaussiana	0,00	0,67	100
$\Delta N_{trimestral}$ Sociologia	3	Não Gaussiana	0,02	0,66	100
ΔN_{anual} Anuidade	1	Não Gaussiana	1025,86	3605,10	100
ΔN_{anual} Anuidade	2	Não Gaussiana	1006,88	3567,70	100
ΔN_{anual} Anuidade	3	Não Gaussiana	995,27	3517,30	100
ΔN_{anual} Desconto	1	Não Gaussiana	0,01	1,06	100
ΔN_{anual} Desconto	2	Não Gaussiana	0,01	0,91	100
ΔN_{anual} Desconto	3	Não Gaussiana	0,01	1,11	100
ΔN_{anual} Qt Faltas	1	Não Gaussiana	1,42	10,07	100
ΔN_{anual} Qt Faltas	2	Não Gaussiana	1,42	9,97	100
ΔN_{anual} Qt Faltas	3	Não Gaussiana	1,51	10,03	100
ΔN_{anual} Artes	1	Não Gaussiana	-0,07	0,86	100
ΔN_{anual} Artes	2	Não Gaussiana	-0,10	0,91	100
ΔN_{anual} Artes	3	Não Gaussiana	-0,10	0,91	100
ΔN_{anual} Biologia	1	Não Gaussiana	-0,01	0,38	100
ΔN_{anual} Biologia	2	Não Gaussiana	-0,02	0,48	100
ΔN_{anual} Biologia	3	Não Gaussiana	-0,01	0,44	100
ΔN_{anual} Ciência	1	Não Gaussiana	-0,12	0,70	100
ΔN_{anual} Ciência	2	Não Gaussiana	-0,12	0,76	100
ΔN_{anual} Ciência	3	Não Gaussiana	-0,10	0,71	100
ΔN_{anual} Ed. Física	1	Não Gaussiana	-0,03	0,74	100
ΔN_{anual} Ed. Física	2	Não Gaussiana	-0,04	0,80	100
ΔN_{anual} Ed. Física	3	Não Gaussiana	-0,02	0,77	100
ΔN_{anual} Filosofia	1	Não Gaussiana	0,00	0,52	100
ΔN_{anual} Filosofia	2	Não Gaussiana	-0,02	0,59	100
ΔN_{anual} Filosofia	3	Não Gaussiana	-0,01	0,58	100
ΔN_{anual} Física	1	Não Gaussiana	-0,01	0,43	100
ΔN_{anual} Física	2	Não Gaussiana	-0,02	0,53	100
ΔN_{anual} Física	3	Não Gaussiana	0,00	0,46	100
ΔN_{anual} Geografia	1	Não Gaussiana	-0,12	0,85	100
ΔN_{anual} Geografia	2	Não Gaussiana	-0,14	0,94	100
ΔN_{anual} Geografia	3	Não Gaussiana	-0,09	0,89	100
ΔN_{anual} História	1	Não Gaussiana	-0,11	0,86	100
ΔN_{anual} História	2	Não Gaussiana	-0,15	0,96	100
ΔN_{anual} História	3	Não Gaussiana	-0,10	0,93	100
ΔN_{anual} Inglês	1	Não Gaussiana	-0,10	0,85	100
ΔN_{anual} Inglês	2	Não Gaussiana	-0,11	1,01	100
ΔN_{anual} Inglês	3	Não Gaussiana	-0,08	0,91	100
ΔN_{anual} Matemática	1	Não Gaussiana	-0,14	0,82	100
ΔN_{anual} Matemática	2	Não Gaussiana	-0,15	0,95	100
ΔN_{anual} Matemática	3	Não Gaussiana	-0,11	0,90	100
ΔN_{anual} Português	1	Não Gaussiana	-0,10	0,67	100
ΔN_{anual} Português	2	Não Gaussiana	-0,11	0,81	100
ΔN_{anual} Português	3	Não Gaussiana	-0,08	0,72	100
ΔN_{anual} Química	1	Não Gaussiana	-0,02	0,45	100
ΔN_{anual} Química	2	Não Gaussiana	-0,03	0,50	100
ΔN_{anual} Química	3	Não Gaussiana	-0,01	0,46	100
ΔN_{anual} Sociologia	1	Não Gaussiana	0,00	0,43	100
ΔN_{anual} Sociologia	2	Não Gaussiana	0,00	0,52	100
ΔN_{anual} Sociologia	3	Não Gaussiana	0,01	0,49	100