

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

**ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR DO ENSINO MÉDIO ATRAVÉS
DE DADOS PÚBLICOS**

por
Fernanda Scovino Machado

Rio de Janeiro
2019

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

**ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR DO ENSINO MÉDIO ATRAVÉS
DE DADOS PÚBLICOS**

"Declaro ser a única autora da presente monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador".

Fernanda Scovino Machado

Orientador: Prof. Flavio Coelho Codeço

**Rio de Janeiro
2019**

FERNANDA SCOVINO MACHADO

**ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR DO ENSINO MÉDIO ATRAVÉS
DE DADOS PÚBLICOS**

“Trabalho de Conclusão apresentado à Escola de Matemática Aplicada como requisito para a obtenção do grau de bacharel em Matemática Aplicada”

Aprovado em ____ de _____ de _____.

Grau atribuído à Monografia: _____.

Professor Orientador: Flavio Coelho Codeço
Escola de Matemática Aplicada – FGV/EMAp
Fundação Getulio Vargas

Agradecimentos

A todas(os) que me inspiraram e ajudaram nessa jornada chamada vida universitária.

Sumário

Agradecimentos	i
Lista de Figuras	iv
Lista de Tabelas	v
1 Introdução	3
1.1 Motivação e Justificativa	3
1.2 Objetivo do Projeto	4
1.3 Estrutura da Monografia	4
2 Revisão de literatura	5
2.1 Combate à evasão escolar	5
2.2 Previsão da evasão escolar	8
2.3 Modelo logístico	8
2.4 Modelos lineares multinível	11
2.4.1 Intercepto aleatório	11
2.4.2 Coeficiente aleatório	12
3 Metodologia	13
3.1 Bases de dados	13
3.1.1 Censo escolar	13
3.1.2 Nível Socioeconômico	14
3.1.3 Geolocalização de escolas, proximidade de comunidades e tiroteios	15
3.2 Pré-processamento	15
3.3 Modelagem	16
3.4 Regressão logística	16
3.4.1 Modelo multinível - Intercepto aleatório	17
3.4.2 Modelo multinível - Coeficiente aleatório	18
3.5 <i>Random Forest</i>	19
4 Resultados	20
4.1 Análise exploratória	20
4.2 Efeito escola	25
4.3 Modelo base: Idade	27
4.4 Intercepto e coeficiente aleatório	28
4.5 <i>Mixed Random Forest</i>	31

<i>SUMÁRIO</i>	iii
5 Conclusões	34
A Pré-processamento das variáveis de matrícula do Censo Escolar	35
B Pré-processamento das variáveis de escola do Censo Escolar	37
Referências Bibliográficas	40

Lista de Figuras

2.1	Motivos da evasão escolar - Projeto Aluno Presente	6
2.2	Motivos da evasão escolar - Projeto Busca Ativa Escolar	7
2.3	Gráfico da função $\text{logit}^{-1}(x)$	10
2.4	Regressão linear para diferentes modelos multinível	12
4.1	Distribuição da idade em função da indicadora de evasão	21
4.2	Distribuição do número de estudantes e taxa de evasão por escola	22
4.3	Distribuição das variáveis numéricas	22
4.4	Efeito escola estimado no modelo nulo	25
4.5	Percentual de cobertura do modelo nulo	26
4.6	Probabilidade estimada pelo modelo base	28
4.7	Percentual de cobertura do modelo base (idade)	28
4.8	Percentual de cobertura do modelo com intercepto aleatório	30
4.9	Probabilidade estimada pelo intercepto aleatório	30
4.10	Intercepto aleatório estimado por escola	31
4.11	Probabilidade estimada pelo <i>mixed random forest</i>	31
4.12	Percentual de cobertura do modelo de <i>mixed random forest</i>	32
4.13	Curvas ROC dos modelos trabalhados	33

Lista de Tabelas

2.1	Resumo dos fatores de risco de evasão escolar no E.M. da literatura	7
3.1	Filtragem gradativa das bases de dados	13
4.1	Distribuição das variáveis a nível aluno em função da indicadora de evasão	20
4.2	Distribuição das variáveis a nível aluno em função da indicadora de evasão	24
4.3	Estimativa do coeficiente de primeiro nível do modelo nulo	25
4.4	Estimativa dos coeficientes do modelo com variáveis de escola	27
4.5	Estimativa dos coeficientes do modelo base	27
4.6	Estimativa dos coeficientes do modelo com intercepto aleatório	29
4.7	Separação das observações no <i>mixed random forest</i>	32
A.1	Pré-processamento das variáveis de alunas(os)	35
B.1	Pré-processamento das variáveis de escola	37

Resumo

A evasão é um dos principais desafios educacionais no Brasil e tem seu auge na transição do ensino fundamental para o ensino médio. Segundo estudo do Insper¹[3], um a cada 4 jovens de 15 a 17 anos interrompem sua trajetória escolar nessa etapa. Diversos programas de combate à evasão surgiram nos últimos anos em diferentes estados e municípios, com foco na busca ativa de jovens fora da escola. O trabalho desenvolvido tem o intuito de explorar modelos para previsão desse fenômeno no universo de estudantes do 1^a série do E.M. de escolas estaduais no Rio de Janeiro com dados do Censo Escolar e geolocalização de escolas e comunidades fluminenses. Para contemplar a estrutura de agrupamento de estudantes em escolas, foram testados modelos de regressão e *random forest* multiníveis, e comparadas suas performances nos dados levantados.

Palavras-chave: evasão, ensino médio, previsão, modelos de aprendizado de máquina

¹Parceria à Fundação Brava, Instituto Unibanco e Instituto Ayrton Senna

Abstract

School dropout is one of the main educational challenges in Brazil and has its peak in the transition from elementary to high school. According to a study by Insper [3], one in every 4 young people aged 15 to 17 interrupt their studies at this stage. Several dropout recovery programs have emerged in recent years in different states and municipalities, focusing on actively seeking out-of-school youth. The aim of this work is to explore models for predicting this phenomenon in the universe of students from the 1st year of the state high schools in Rio de Janeiro with data from the School Census, aggregated to schools and favelas geolocation. To contemplate the grouping structure of students in schools, multilevel regression and random forest models were tested, and their performances compared in the surveyed data.

Keywords: dropout, high school, prediction, machine learning models

Capítulo 1

Introdução

1.1 Motivação e Justificativa

A evasão escolar é um problema grave na educação. Segundo pesquisa da ONG Todos pela Educação com os dados do PnadC¹, em 2018, mais de 1 milhão de jovens de até 19 anos (36,5% do total) ainda não terminaram o ensino médio (E.M.). Dentre esses, 62% não frequentam mais a escola, pouco mais da metade deixando o ambiente escolar ainda no ensino fundamental[15]. Isso representa um gargalo da administração pública na sua missão de universalização do atendimento escolar, prevista no Art. 2º II do Plano Nacional de Educação (PNE), Lei nº. 13.005, de 25 de junho de 2014 [5]; mais especificamente referente à meta 3 de: "universalizar, até 2016, o atendimento escolar para toda a população de 15 a 17 anos e elevar, até o final do período de vigência deste PNE [até 2024], a taxa líquida de matrículas no ensino médio para 85%."

O abandono dos estudos também se caracteriza como um forte fator para o aumento de desigualdade socioeconômica. Diversos estudos comprovam que o retorno da educação no Brasil não só é muito elevado [2], como também cresce de acordo com o nível de escolaridade [14] [9], ou seja, a variação salarial média de um jovem com diploma do ensino fundamental (E.F.) para um diploma do ensino médio é menor que quando comparada à variação do ensino médio para a graduação. Isso se concretiza num horizonte reduzido de oportunidades: um jovem sem o diploma do E.M. não consegue continuar os estudos e por consequência não tem acesso a empregos com maior renda, o que afeta principalmente os jovens mais vulneráveis.

No estado do Rio de Janeiro (RJ), apenas 75,3% dos jovens de até 16 anos concluíram o ensino fundamental (E.F.) e 62,7% dos jovens de até 19 anos concluíram o E.M. [15]. Nos últimos 6 anos (2012 a 2018), a taxa de conclusão do E.F.² aumentou em 9,3%, com a 11ª maior variação entre os estados, porém a taxa de conclusão do E.M.³ teve um aumento de somente 5,9%, colocando o estado na penúltima posição⁴. Isso nos alerta para um estudo mais aprofundado do cenário do E.M. no estado, e quais os obstáculos enfrentados por estudantes nessa mudança de etapa.

Diversas iniciativas já foram e ainda são realizadas no combate à evasão, com um desta-

¹Pesquisa Nacional por Amostra de Domicílios Contínua

²Variação do número de jovens até 16 anos que concluíram o E.F.

³Variação do número de jovens até 19 anos que concluíram o E.M.

⁴Cálculo a partir dos dados das Taxas de Transição do INEP

que àquelas que se utilizam da busca ativa de estudantes fora da escola e acompanhamento desta(e) na sua reinserção ao ambiente escolar. Esses programas muitas vezes agregam órgãos da educação, saúde, psicólogas(os) e voluntárias(os) que trabalham junto à família da(o) estudante. A possibilidade de identificação do risco de evasão seria um auxílio no mapeamento de jovens para participação desses programas e no desenho de políticas públicas locais evitando que sua trajetória seja interrompida.

1.2 Objetivo do Projeto

Tendo em vista o exposto acima, este projeto tem por objetivo identificar estudantes em risco de evasão escolar na 1ª série do ensino médio das escolas estaduais do estado do Rio de Janeiro, a partir da análise dos dados públicos de educação. Buscamos estudar características de estudantes e escolas com base na literatura, verificando se estas se apresentam como bons preditores para o fenômeno da evasão.

1.3 Estrutura da Monografia

O trabalho está dividido em 5 capítulos: Introdução, Revisão de literatura, Metodologia, Resultados, Conclusão. Este primeiro capítulo apresentou uma breve introdução ao projeto a ser trabalhado na monografia, assim como o contexto no qual se insere. O próximo capítulo discute o estado atual do problema e iniciativas locais de combate à evasão escolar, estudos realizados para predição e caracterização do fenômeno, e conceitua-se os modelos que serão utilizados.

O Capítulo 3 descreve em detalhes as base e o pré-processamento dos dados, a modelagem do problema e principais hipóteses para o desenvolvimento do projeto. No Capítulo 4 é apresentada a análise exploratória das variáveis selecionadas, os resultados dos modelos trabalhados e a comparação de suas performances. Por fim, no Capítulo 5 tem-se a conclusão da monografia, suas limitações e possíveis trabalhos futuros para o projeto.

Capítulo 2

Revisão de literatura

2.1 Combate à evasão escolar

No âmbito da administração pública, visto que o E.M. é de responsabilidade dos estados¹, existem políticas estaduais de prevenção da evasão escolar no E.M. que buscam combater fatores relacionados à falta de engajamento, violência e trabalho infantil, porém boa parte foi descontinuada nos últimos anos². Um extenso estudo do Insper (2017)[3] sobre as políticas de combate à evasão e o fracasso escolar no Brasil apontam as principais iniciativas tomadas no estado nos últimos anos. Um dos exemplos é o Programa Renda Melhor Jovem, foi implantado em 2011 e suspenso em 2014.³ Nele, a(o) estudante cuja família já participasse de algum programa de transferência de renda (Renda Melhor ou Cartão Família Carioca) poderia receber um valor anual de R\$700 (1ª série) a R\$1,2 mil (4ª série - Ensino profissionalizante) a cada série concluído no E.M.

Além dessas ações, são desenvolvidas também políticas de tratamento da evasão com a reinserção de estudantes fora da escola, como o Aluno Presente e o Busca Ativa Escolar. O projeto Aluno Presente foi uma iniciativa da Associação Cidade Escola Aprendiz em parceria com a Secretaria Municipal de Educação (SME) do Rio de Janeiro e com apoio da Fundação *Education Above All* que mapeou crianças e jovens de 6 a 14 anos fora da escola no município. O projeto contou com mais de 1,5 mil instituições parceiras e, segundo dados oficiais do mesmo, identificou 23.753 crianças e adolescentes e reinseriu 22.131 delas em unidades escolares de 2013 a 2016, chegando à meta estabelecida de alcançar 90% das crianças e adolescentes nesta situação na cidade do Rio de Janeiro.[1]. A identificação de estudantes fora da escola foi realizada através da busca ativa em 5 regiões da cidade definidas pela SME, com 3 diferentes estratégias: busca de campo voluntária, consulta a listas oficiais da SME de estudantes que estavam sem frequentar a escola e indicações de casos por instituições locais.⁴ Isso permitiu com que fossem mapeados os principais motivos de evasão do E.F. no município, apresentados na Figura 2.1.

Embora seja um projeto focado no E.F., a metodologia do Aluno Presente posteriormente serviu de inspiração para a criação da estratégia Busca Ativa Escolar, da iniciativa Fora da Escola Não Pode! da UNICEF, que abrange desde o E.F. ao E.M. com a adesão de estados

¹Conforme termos do art. 211 §3º da Constituição Federal

²Práticas por localidade e fatores determinantes em: <http://gesta.org.br/tema/engajamento-escolar/#painel>

³Disponível em: <http://www.rendamelhorjovem.rj.gov.br/index.php/rmjovem/>

⁴Mais detalhes em: <https://www.alunopresente.org.br/estrategias/busca-ativa/>

MOTIVO PRINCIPAL DECLARADO PARA A CRIANÇA ESTAR FORA DA ESCOLA	%
Mudança de domicílio, viagem ou deslocamentos frequentes	16,4
Aspectos relacionados a questões familiares ou de vulnerabilidade social	15,1
Fatos que dificultam o deslocamento / acesso à escola	14,3
Desinteresse ou desistência da criança pela escolarização	12,9
Criança com demanda de saúde ou necessidades educacionais especiais	11,1
Criança com conflito interpessoal ou dificuldade de adaptação no ambiente escolar	7,2
Insuficiências relacionadas a prestação do serviço educacional	6,2
Outros	9,9
Motivo desconhecido / não informado	6,9
TOTAL	100

Figura 2.1: Motivos da evasão escolar no E.F. mapeados pelo projeto Aluno Presente no município do Rio de Janeiro

e municípios na plataforma. A Busca Ativa Escolar também trabalha na identificação de crianças e jovens fora da escola com estratégias semelhantes. Uma vez identificado um caso, é criado um alerta na plataforma com dados cadastrais coletados e o motivo da evasão, e um técnico verificador recebe o alerta para validá-lo e realizar visitas às famílias mapeadas. O gestor político do projeto atribui critérios de priorização para os motivos reportados de forma a atender à realidade do estado ou município. Posteriormente, as(os) estudantes são acompanhados na sua reinserção no ambiente escolar ao longo de 1 ano antes de concluir o caso.

Além da busca de campo, a plataforma se utiliza dos relatórios de evasão do Educacenso, que identificam como estudantes evadidas(os) aqueles registrados no Censo Escolar no ano anterior e que não aparecem no seguinte. O Censo Escolar é o principal instrumento de coleta de informações da educação básica e a mais importante pesquisa estatística educacional brasileira⁵, preenchido anualmente pelas secretarias estaduais e municipais com informações sobre estudantes, turmas, escolas e professoras(es). Mais de 20% dos alertas são gerados a partir desse relatório, como mostra a Figura 2.2, que pode servir como uma boa ferramenta para uma primeira análise dos perfis de estudantes fora da escola - mas ainda não é utilizado de forma estruturada para a previsão da evasão escolar com o mapeamento de risco de estudantes.

O estudo desses casos levou à identificação da relação direta entre a distorção idade-série⁶ e a evasão escolar nos estados e municípios participantes. Assim, foi lançada uma segunda estratégia da iniciativa Fora da Escola Não Pode!, a Trajetórias de Sucesso Escolar, para apoiar estados e municípios na definição, implementação e avaliação de políticas e ações de enfrentamento da distorção idade-série e superação do fracasso escolar [17].

⁵Disponível em: <http://portal.inep.gov.br/censo-escolar>

⁶Estudante com 2 ou mais anos acima da idade adequada para a etapa de ensino em que se encontra



Figura 2.2: Motivos da evasão escolar no E.F. e E.M. mapeados pelo Busca Ativa Escolar no estado do Rio de Janeiro

Com base nas iniciativas apresentadas e pesquisas na área (discutidas nas próximas sessões), resumimos os principais fatores da literatura em 4 temáticas na Tabela 2.1 abaixo. Sabemos que os fatores de diferentes temas não são independentes entre si, mas sim agem de forma concomitante. Optamos por essa separação dada a possibilidade de mapeamento de cada fator com diferentes conjuntos de dados - com o a tabela de Escolas do Censo Escolar, por exemplo, conseguimos trabalhar todos os dados referentes à infraestrutura. Fatores mais específicos a nível aluno, como trabalho, gravidez e déficit de aprendizagem, não podem ser mapeados com as bases disponíveis hoje pelo INEP (ou não é possível o cruzamento destas com outras para identificação de estudantes).

Socioeconômico	Trabalho, pobreza (1) , escolaridade dos pais, Bolsa Familia (2)
Socioemocional	Gravidez, violência (1)
Infraestrutura	Acesso limitado (1), laboratórios (4), localização da escola (3, 4)
Aprendizado	Déficit de aprendizagem (1), distorção idade-serie (2, 3) , repetência (2), turno de estudo (3), etapa de ensino (4)

Tabela 2.1: Resumo dos fatores de risco de evasão escolar no E.M. da literatura. Em destaque, os fatores que acreditamos serem contemplados com os dados utilizados no presente estudo. A numeração refere-se às citações de fatores apresentados em (1) Insper, 2017; (2) Shirasu et. al (2016); (3) Bezerra et al. (2016); (4) Calixto et. al (2017).

2.2 Previsão da evasão escolar

Dentre os trabalhos nacionais de previsão da evasão no ensino médio (E.M.), muitos se utilizam do Censo Escolar, que abrange informações referentes a estudantes, docentes e unidades escolares. Nesses estudos não são incluídos diretamente notas de avaliações de desempenho, como o Saeb ou Prova Brasil, visto que os dados de proeficiência disponibilizados a nível aluno não são comparáveis com os dados do Censo (não existe um identificador comum entre ambos).

Bezerra et al. (2016) [4] desenvolveram um estudo acerca das variáveis de impacto da evasão escolar no último ano do ensino fundamental (E.F. - 9º ano) no estado de Pernambuco nos anos de 2011 e 2012 com os dados do censo escolar. O trabalho comparou a performance e as variáveis explicativas de três diferentes modelos: árvore de decisão, indução de regra e regressão logística. Os resultados foram consistentes entre os modelos, indicando uma maior importância de fatores como idade (primeiro nível de separação da árvore), turno de estudo e região geográfica.

Mais recente, Calixto et. al (2017) [6] buscaram estimar e comparar preditores de evasão nos anos de 2014, 2015 e 2016 com estudantes do estado do Ceará e Sergipe, respectivamente os estados com maior e menor proeficiência no IDEB⁷. A partir dos dados do censo escolar de diferentes etapas de ensino (não só E.M.), testaram-se modelos de indução de regra e regressão logística. Os autores reportam uma acurácia de aproximadamente 87% da regressão para ambos os estados, destacando também dentre os fatores de maior impacto a idade da(o) estudante, a etapa de ensino (maior probabilidade de evasão no E.M.), modalidade de ensino (E.M. Regular/ Profissionalizante), existência de laboratórios e localização da escola. Em geral, os estados tiveram fatores semelhantes, embora no estado do Ceará a raça negra também teve influência positiva na probabilidade de evasão - não foi observado nenhum efeito dessa variável para o Sergipe.

Vale destacar também o papel de estudos com dados socioemocionais de estudantes, como Caluz (2018) [7], na tentativa de "explorar a lacuna existente na atual literatura, que é insuficiente para explicar a estagnação da desigualdade de conclusão do ensino médio entre famílias mais ricas e mais pobres". O estudo, realizado no interior de São Paulo, indica que características como aceitação e pertencimento da(o) estudante ao ambiente escolar possuem um papel preditivo na permanência do mesmo na escola.

2.3 Modelo logístico

A regressão logística é um modelo linear largamente utilizado na previsão de respostas binárias por ser um dos mais simples e ter uma boa interpretabilidade. No modelo, a probabilidade p_i de realização de um determinado fenômeno para a observação i - representada pela variável resposta $y_i \in \{0, 1\}$ - é dada por:

$$Pr(y_i = 1) = \hat{p}_i \quad (2.1)$$

$$\text{logit}(\hat{p}_i) = \beta_0 + \beta^T X_i \quad (2.2)$$

⁷Índice de Desenvolvimento Nacional da Educação Básica, construído a partir do Saeb - avaliação nacional bianual realizada nos anos iniciais e finais do Fundamental

sendo $\tau_i = \beta_0 + \beta^T X_i$ o *preditor linear* do fenômeno modelado, β_0 o coeficiente independente, $\beta = [\beta_1, \dots, \beta_k, \dots, \beta_K]^T$ o vetor dos coeficientes estimados pelo modelo para cada variável explicativa $x_{k,i} \in X_i$.

Assumimos que a distribuição de probabilidade da variável aleatória de realização do fenômeno segue uma Bernoulli, isto é:

$$Pr(y = 1) = p \quad (2.3)$$

$$Pr(y = 0) = 1 - p \quad (2.4)$$

A estimação \hat{p}_i gerada da regressão logística busca aproximar y_i , modelando um valor p_i para cada observação dado seu preditor linear τ_i . Intuitivamente, o que queremos da função $p_i(\tau_i)$ é que ela suba rapidamente de 0 para 1 quando $p_i \approx p$ - ou seja, que separe os sucessos e fracassos da realização do fenômeno razoavelmente bem.

Dada a probabilidade p_i de sucesso e $1 - p_i$ de fracasso, a proporção de sucessos em relação a fracassos, $\frac{p_i}{1-p_i}$, traduz a *chance* da realização do fenômeno. A função *logit* é definida então como o "*log das chances de sucesso*":

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots \beta_K X_{K,i} \quad (2.5)$$

Essa função é utilizada por conseguir aproximar linearmente os valores $\log\left(\frac{p}{1-p}\right)$ no intervalo $p \in [0, 1] \rightarrow \log\left(\frac{p}{1-p}\right) \in (-\infty, \infty)$ a partir dos preditores lineares τ_i . Podemos manipular a equação 2.5 para obter p em função de τ da seguinte forma:

$$\log\left(\frac{p_i}{1-p_i}\right) = \tau_i \quad (2.6)$$

$$\frac{p_i}{1-p_i} = e^{\tau_i} \quad (2.7)$$

$$p_i(1 + e^{\tau_i}) = e^{\tau_i} \quad (2.8)$$

$$p_i = \frac{1}{1 + e^{-\tau_i}} \quad (2.9)$$

Logo, para cada valor τ_i temos um p_i associado, modelado como uma sigmoide (Figura 2.3), na qual $\lim_{\tau \rightarrow -\infty} = 0$ e $\lim_{\tau \rightarrow \infty} = 1$.

Por construção da função $\text{logit}^{-1}(x) = y$, as mudanças na escala *logit* são suprimidas nas pontas da distribuição, garantindo que a mesma se mantenha na escala de 0 a 1 [16]. A regressão logística estima, então, os parâmetros β_k na equação:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots \beta_K X_{K,i} \quad (2.10)$$

Os coeficientes (β_k) também são chamados de *efeitos fixos* do modelo - pois são constantes para todas as observações. Para a correta interpretação do modelo, é necessário termos como premissa que as variáveis resposta y_i são independentes dada sua probabilidade [10].

A estimação dos parâmetros deve ser interpretada exponencialmente como o grau de mudança da *razão das chances* de realização do fenômeno. Por exemplo, se $\beta_k = 0.2$, isso significa que uma unidade de diferença em x_k corresponde a uma mudança proporcional a

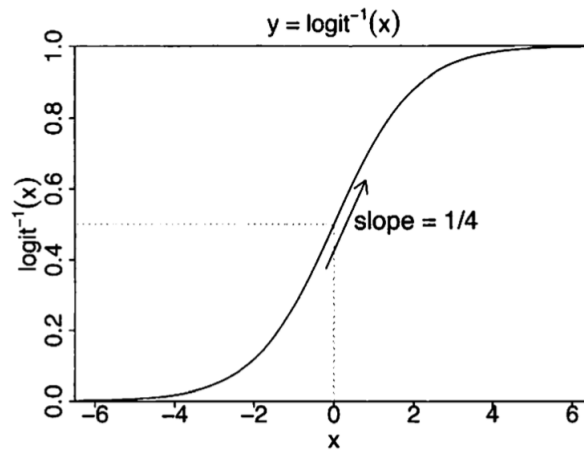


Figura 2.3: Função $\text{logit}^{-1}(x)$: transformação dos preditores lineares para as probabilidades estimadas na regressão logística (Fonte: Gelman e Hill (2006) [10])

$\exp^{0.2} = 1.22$, ou seja, as chances de evasão poderiam ir de 1:1 para 1.22:1, o que pode não ser tão intuitivo.

Para entendermos a relação dos coeficientes com mudanças na probabilidade, é necessário notar que a mudança esperada em y dada uma mudança fixa em x não é constante [10] (como mostra a Figura 2.3). Mais especificamente, a maior variação que podemos observar em y é obtida entorno de $x = 0.5 \rightarrow \beta X = 0$, ponto no qual a derivada de logit^{-1} toma seu maior valor, dado pela equação abaixo.

$$\text{logit}^{-1}(\beta X) = \frac{1}{1 + \exp \beta X} \quad (2.11)$$

$$(\text{logit}^{-1}(\beta X))' = \frac{\beta \exp X}{(1 + \exp X)^2} \quad (2.12)$$

$$(\text{logit}^{-1}(0))' = \frac{\beta}{4} \quad (2.13)$$

Logo, é razoável tomarmos $\frac{\beta}{4}$ como limite superior para a mudança da probabilidade em função do coeficiente β . Para valores próximo a 0.5, podemos ainda afirmar que esse limite é um bom aproximador da variação esperada da probabilidade. No exemplo anterior ($\beta = 0.2$), supondo $\text{logit}(\hat{p}_i) = -0.3 + 0.2x$, temos que a variação esperada da probabilidade em relação a x é de $\approx \frac{0.2}{4} = 0.05$, ou seja, espera-se uma diferença unitária em x cause aumento em até 5% a probabilidade.

Já a previsão da variável resposta, \hat{y} , pode ser interpretada em função de um limite δ que determinamos para a separação da probabilidade estimada: $\hat{y}_i = \{1, \text{ se } \hat{p}_i \geq \delta, 0 \text{ c.c.}\}$. Essa configuração é chamada de *variável latente*, pois assumimos existir um δ não observado. Algumas das principais vantagens dessa formulação são uma maior margem de interpretação dos resultados e sua utilidade computacionalmente para calcularmos métricas acerca do modelo (mais detalhes na seção 3.4.1).

2.4 Modelos lineares multinível

Estudos econométricos das causas de evasão escolar são bastante frequentes na pesquisa de economia da educação. Esses estudos contribuem no presente trabalho, principalmente, com a utilização de modelos multiníveis que incorporam características de agrupamento de estudantes (turmas e/ou escolas). É o caso de Shirasu et. al (2016) [16], que utilizam os dados do Sistema de Avaliação Permanente da Educação Básica do Ceará (SPAECE) para a determinação dos fatores de influência na evasão e repetência escolar no E.M. de 2008 a 2011. O SPAECE é uma avaliação diagnóstica local aplicada no E.F. (5º e 9º ano) e E.M. Por abranger todos os anos do E.M., os resultados permitiram o acompanhamento do desempenho de estudantes e a construção da variável de evasão com base naquelas(es) que participaram em 2008 e não foram encontrados em nenhum ano posterior. Combinados a dados de infraestrutura das escolas do Censo Escolar, foi construído um modelo de regressão multinível, validado na literatura levantada pelos autores, que buscou medir o impacto de fatores como o desinteresse dos jovens (traduzido na indicadora se a(o) estudante fazia o dever de casa), o nível de escolaridade dos pais, a participação no programa Bolsa Família, repetência e distorção idade-série, e também sua relação com a escola e questões de infraestrutura. O estudo alerta que estudantes com alguma repetência anterior tem quase o dobro das chances de evadir[16], e outras variáveis como o interesse e nível de escolaridade dos pais também se mostraram significativas.

Os modelos multiníveis ou hierárquicos são empregados em diferentes contextos, podendo ser construídos tanto para fins de previsão quanto inferência causal e modelos descritivos [10]. O uso desse tipo de modelagem parte da premissa de existir algum nível de agrupamento dos dados a serem trabalhado (por exemplo, na previsão de notas de estudantes de diferentes escolas). A estrutura hierárquica dos dados quebra a premissa de modelos lineares padrão (nível único) de que os erros observados são independentes entre si.

A regressão multinível pode ser pensada como um método que reside entre os dois extremos: a exclusão de um preditor que contenha categorias dos agrupamentos do modelo (*complete pooling*) - por exemplo, desconsiderar o agrupamento das escolas como variáveis no modelo -, ou a estimativa de modelos separados em cada nível de agrupamento (*no pooling*) - rodar um modelo independente para cada escola. O problema dessas abordagens é que, no primeiro caso, pode levar a inconsistência na análise da variação dos coeficientes dado que os fatores não são independentes entre si, e, no segundo caso, estamos desconsiderando possíveis fenômenos que perpassam os grupos, suprimindo variações importantes para análise [10]. A incorporação do fator de agrupamento pode ser feita de duas formas: construção de um intercepto e/ou coeficiente aleatório.

2.4.1 Intercepto aleatório

Nesse modelo, permitimos que o intercepto do modelo linear ($\beta_{0,j}$) varie para diferentes grupos j . Os interceptos são estimados no segundo nível do modelo para cada grupo $j \in J$, de acordo com um conjunto de características $W_{k,j}$ do grupo (por exemplo, dados de infraestrutura de escolas). A fórmula é dada por:

$$y_i = \alpha_{j[i]} + \beta_1 X_{1,i} + \dots + \beta_{M,i} X_{M,i}, \quad i = 1, \dots, n \quad (2.14)$$

$$\alpha_j = \alpha_0 + \sum_{k=1}^K \gamma_k W_{k,j} + r_{0,j}, \quad j = 1, \dots, J \quad (2.15)$$

sendo $W_j = [W_{j,1}, \dots, W_{j,K}]$ o vetor das K variáveis referentes ao grupo j , γ_k seus respectivos efeitos estimados (no segundo nível), α_0 o intercepto fixo e $r_{0,j}$ o resíduo da regressão referente ao agrupamento da escola j .

Esse modelo consegue capturar o que chamamos de *efeito de grupo* através do intercepto aleatório, porém não é feita distinção do efeito das variáveis de estudantes entre escolas (ex: idade na escola a pode ser um fator com maior impacto para evasão do que na escola b , ou até mesmo ter um efeito inverso). Para introduzir esta variação, precisamos de outra configuração de modelo multinível - chamado coeficiente aleatório.

2.4.2 Coeficiente aleatório

De forma semelhante ao anterior, o modelo de coeficiente aleatório estima coeficientes num segundo nível para cada grupo j - porém nesse caso ao invés de variarmos o intercepto, iremos variar coeficientes $\beta_{m,j}$ referentes às variáveis dos indivíduos no primeiro nível. A expressão do modelo é dada por:

$$y_i = \alpha + \beta_{1,j[i]} X_{1,i} + \dots + \beta_{M,j} X_{M,i}, \quad i = 1, \dots, n \quad (2.16)$$

$$\beta_{m,j} = \beta_0 + \sum_{k=1}^K \gamma_k W_{k,j} + r_{0,j}, \quad j = 1, \dots, J \quad (2.17)$$

sendo $\beta_{m,j}$ o coeficiente da variável a nível indivíduo X_m para o grupo j .

Esse modelo nos permite medir a variação de fatores individuais atribuída ao agrupamento, às vezes indicando efeitos inversos para diferentes grupos. Os dois modelos - intercepto e coeficiente aleatórios - podem também ser combinados. Gelman e Hill (2006) [10] ilustram as diferentes retas geradas por cada tipo de regressão linear multinível (Figura 2.4).

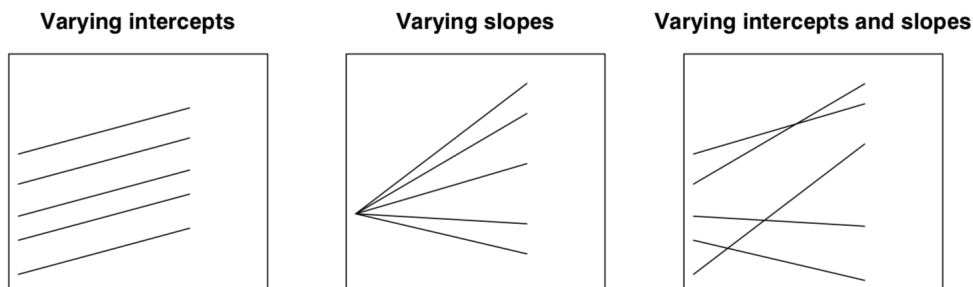


Figura 2.4: Modelos de regressão linear com (a) intercepto aleatório ($y = \alpha_j + \beta x$); (b) coeficiente aleatório ($y = \alpha + \beta_j x$); (c) intercepto e coeficiente aleatório ($y = \alpha_j + \beta_j x$). Para facilitar a visualização, o modelo possui somente uma variável preditora x .

Capítulo 3

Metodologia

3.1 Bases de dados

Para o desenvolvimento do estudo, utilizaremos 4 (quatro) bases de dados principais: Censo escolar e Nível Socioeconômico (INSE), do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep); geolocalização de escolas e comunidades, do MP em Mapas; e tiroteios registrados no aplicativo Fogo Cruzado no estado no Rio de Janeiro em 2016. A interseção dessas bases define nosso escopo de trabalho no projeto, que está resumida na tabela abaixo - os dados de tiroteio entraram como indicadores, logo não restringem as escolas da base. Reduzimos ainda o número de estudantes a partir da identificação de idades muito discrepantes na cauda da distribuição (apresentada no próximo capítulo), definindo um universo de 146.214 estudantes da 1ª série do E.M. em 837 escolas estaduais no Rio de Janeiro.

Filtro	Número de estudantes	Número de escolas	Evadidas(os) (%)
Escolas com Censo 2016 e 2017	182838	1097	33.85
Escolas com INSE	149234	849	31.88
Escolas com geolocalização	147267	837	31.95

Tabela 3.1: Filtragem gradativa das bases de dados

3.1.1 Censo escolar

Os dados do Censo Escolar¹ são coletados anualmente e disponibilizados na página do Inep. A partir dele, é possível obter informações detalhadas de três entidades:

- Matrículas: Características físicas de estudantes², etapa de ensino, meios de transporte escolar;
- Turmas: quantidade de matrículas, horários, dias da semana, disciplinas, atividades complementares;

¹Disponível em: <http://inep.gov.br/web/guest/censo-escolar>

²Sexo, cor/raça, deficiência física

- Professoras(es): Características físicas (sexo, cor/raça, deficiência física), escolaridade, disciplinas ministradas, tipo de contratação;
- Escolas: número de estudantes, salas, docentes, dependências existentes (refeitório, biblioteca, laboratórios);

Essas são apenas algumas das muitas informações: as tabelas de matrículas, turmas, docentes e escolas possuem, respectivamente, 92, 88, 132 e 168 variáveis, embora muitas estejam sobrepostas entre tabelas. Esses dados vem agregados por regiões em arquivos comprimidos (.zip) dentro de um arquivo maior, contendo todas as regiões brasileiras. Existe aí uma dificuldade inicial de utilização dos dados devido a necessidade de alto processamento para sua extração (somente a região sudeste compreende mais de 4GB de dados). Foi utilizado para a primeira filtragem dos dados o sistema de armazenamento e banco de dados do *Google Cloud Device*³, com as edições do Censo de 2016 e 2017.

A escolha de análise dos anos de 2016 a 2017 deve-se (i) a necessidade do cálculo da variável indicadora da evasão, ou seja, ao menos 2 anos letivos; (ii) ao fato de uma mudança no sistema de identificação dos anos a partir de 2018, o que não torna possível a comparação com os demais anos. Essa escolha, agregada à restrição do estado do Rio de Janeiro, faz que com que o cálculo da evasão possa estar superestimado pois não estamos separando casos de mudança para outro estado e também é possível que a(o) estudante tenha retornado os estudos nos anos seguintes.

3.1.2 Nível Socioeconômico

O indicador de nível socioeconômico (INSE) é também disponibilizado pelo (INEP)⁴, mas somente a nível escola. Ele sintetiza a relação entre a escolaridade, a ocupação e a renda das famílias com base nas respostas de estudantes aos formulários socioeconômicos do Saeb⁵ e do ENEM.

O nível socioeconômico é calculado a partir de uma amostra de estudantes, atribuindo-se para cada estudante um valor que segue uma distribuição gaussiana de média 50 e desvio padrão 10, variando conforme as características mencionadas anteriormente (mais detalhes estão disponíveis nas notas técnicas ⁶). O valor do INSE da escola é calculado a partir da média aritmética do nível de suas(seus) estudantes, e então caracteriza-se a mesma como pertencente a um dos 7 grupos de escola, determinados a partir de uma clusterização das mesmas, do nível mais baixo ao mais alto.

³Plataforma de serviço de computação em nuvem do *Google*

⁴Disponível em: <http://portal.inep.gov.br/web/guest/indicadores-educacionais>

⁵Sistema de Avaliação Educacional Brasileiro, prova de avaliação do 6º e 9º ano

⁶Notas técnicas - Inse (2015): <http://portal.inep.gov.br/web/guest/indicadores-educacionais>

3.1.3 Geolocalização de escolas, proximidade de comunidades e tiroteios

O Ministério Público do Rio de Janeiro disponibiliza, através do MP em Mapas⁷, dados de geolocalização agregados de diversas fontes (IBGE, IPP, SABREN/Rio⁸). Utilizamos o conjunto de dados dos limites de comunidades fluminenses e a localização de escolas para estudar também possíveis efeitos na evasão.

As comunidades são delimitadas por polígonos e as escolas são identificadas como pontos no mapa do Rio de Janeiro. A partir desses dados foram criadas duas indicadoras ($I_j^{(r)}$): presença de uma ou mais comunidades no raio de 1km e 5km da escola. O cálculo foi feito centrando-se um círculo de raio r em cada escola (C_j^r - aproximadamente 1km e 5km) e verificando a interseção do círculo com uma comunidade (P):

$$I_j^{(r)} = \begin{cases} 1, & \text{se } \exists P \in \Omega : \{P \cap C_j^r\} \neq \emptyset \\ 0, & \text{c.c.} \end{cases} \quad (3.1)$$

sendo $\Omega = \{\text{conjunto de polígonos de comunidades}\}$ e $C_j^r = \{x \in \mathbb{R}^2 : |x - x_j| \leq r\}$

Da mesma forma, criamos duas indicadoras de tiroteios em 2016 no raio de 1km ou 5km ao redor da escola com os dados do Fogo Cruzado. O aplicativo está no ar desde 2015 e coleta informações de tiroteios via notificação de usuários e outras fontes, como canais de autoridade policiais, validando e publicando a informação. Os dados estruturados foram lançados esse ano e estão disponíveis em sua API⁹.

3.2 Pré-processamento

A tabela de matrículas nos fornece o registro de 182.838 estudantes da 1ª série do ensino médio no estado do Rio de Janeiro. Foi utilizado como identificador único de estudantes o Código INEP, após verificarmos a inconsistência dos identificadores de matrícula ao longo dos anos - apenas 7% se mantinha de 2016 para 2017 em contraponto a 68% dos códigos de identificação única do INEP. Além disso, determinamos os critérios para exclusão de variáveis:

1. Baixo nível de preenchimento (35 variáveis): a variável possui menos de 30% de valores observados
2. Baixa variação (40 variáveis): um único valor da variável concentra mais de 80% das observações
3. Informação duplicada (8 variáveis): já aparece em outro campo ou não é relevante, e informações referentes a escolas que aparecem na outra base

Dentre as 92 variáveis que compõem a tabela de matrículas, foram selecionadas *a priori* apenas 9 variáveis - a tabela em detalhes segue no apêndice A. Quanto à tabela de escolas, de

⁷Disponível em: <https://github.com/MinisterioPublicoRJ/inloco>

⁸Sistema de Assentamento de Baixa Renda

⁹Disponível em: <https://fogocruzado.org.br/sobre-a-api/>

um total de 168 variáveis, foram filtradas 23 para análise de acordo com os critérios abaixo (detalhes também no apêndice B). Além dessas, foram adicionadas 5 variáveis referentes a: nível socioeconômico (valor), indicadora proximidade a comunidades no raio de 1 e 5km, e indicadora de tiroteios registrados em 2016 no raio de 1 e 5km; totalizando 28 variáveis de escola.

1. Baixo nível de preenchimento (13 variáveis): a variável possui menos de 30% de valores observados
2. Baixa variação (111 variáveis): um único valor da variável concentra mais de 80% das observações
3. Informação duplicada (10 variáveis): já aparece em outro campo ou não é relevante, e informações referentes a escolas que aparecem na outra base
4. Alta correlação (10 variáveis): maior que 0.8 com outras variáveis

3.3 Modelagem

A unidade de análise será i , que refere-se a uma estudante matriculada(o) na 1ª série do ensino médio da rede estadual em 2016 (t). A variável indicadora de evasão y_i dessa(e) estudante é definida como:

$$y_i = \begin{cases} 1, & \text{se } i \notin I^{t+1} \\ 0, & \text{c.c.} \end{cases} \quad (3.2)$$

sendo $I^{t+1} = \{\text{conjuntos de estudantes matriculados em 2017 } (t + 1)\}$.

Quanto ao nível de agrupamento de estudantes, iremos nos referir por $j[i]$ a escola j a qual a(o) estudante i pertence. Essa nomenclatura será usada nos modelos multinível testados no trabalho.

3.4 Regressão logística

Conforme descrito na literatura (Seção 2.3), utilizaremos o modelo de regressão logística para estimar as probabilidades de evasão. No modelo padrão (um único nível), a probabilidade para a(o) estudante i , \hat{p}_i será modelada com os preditores a nível aluno (sexo, idade, etc). O modelo é dado por:

$$\text{logit}(\hat{p}_i) = \alpha + \sum_{m=1}^M \beta_{m,i} X_{m,i} + r_i, \quad i = 1, \dots, n \quad (3.3)$$

sendo $X_i = [X_{1,i}, \dots, X_{i,M}]$ o vetor das m variáveis referentes a(o) estudante i , β_m seus respectivos efeitos fixos estimados, α o intercepto (coeficiente independente do modelo) e r_i o resíduo da regressão para a(o) estudante i .

Nos modelos de regressão logística multiníveis, passamos a considerar a escola j a qual a(o) estudante i pertence, $j[i]$, como um nível de agregação de estudantes, que passa a variar alguns dos coeficientes por escola.

3.4.1 Modelo multinível - Intercepto aleatório

Nesse modelo, os interceptos ($\beta_{0,j}$) são estimados num segundo nível para cada escola $j \in J = [\text{conjunto das escolas}]$, de acordo com um conjunto de características $W_{k,j}$ da escola, i.e., passam a ser modelados como *efeitos aleatórios*. A expressão do modelo é dada por:

$$\text{logit}(\hat{p}_i) = \alpha_{j[i]} + \beta_1 X_{1,i} + \dots + \beta_M X_{M,i}, \quad i = 1, \dots, n \quad (3.4)$$

$$\alpha_j = \alpha_0 + \sum_{k=1}^K \gamma_k W_{k,j} + r_{0,j}, \quad j = 1, \dots, J \quad (3.5)$$

sendo $W_j = [W_{j,1}, \dots, W_{j,K}]$ o vetor das K variáveis referentes à escola j (por exemplo: localização, nível socioeconômico, etc), γ_k seus respectivos efeitos fixos estimados (no segundo nível), α_0 o intercepto fixo e $r_{0,j}$ o resíduo da regressão referente ao agrupamento da escola j .

A utilização de modelos multiníveis é atribuída ao fato de acreditarmos que haja uma interdependência dos indivíduos devido a um determinado agrupamento dos mesmos (no caso, estudantes em escolas). Comumente na literatura essa dependência é calculada participando a variância residual para cada grupo de indivíduos [11], para verificar se a variância entre indivíduos de um mesmo grupo (*between*) é menor do que a variância entre todos (*within*). Construímos então o que é chamado de *modelo nulo* [16] considerando somente os agrupamentos dos indivíduos, como na equação 3.7.

$$y = \alpha_{0,j[i]} \quad (3.6)$$

$$\alpha_{0,j} = \alpha_0 + r_{0,j}, \quad j = 1, \dots, J \quad (3.7)$$

Esse modelo é utilizado para o cálculo do *coeficiente de partição da variância (VPC)*, que mensura qual a proporção da variância residual pode ser explicada por diferenças entre grupos (equação 3.8).

$$VPC = \frac{\sigma_{r0}^2}{\sigma_{r0}^2 + \sigma_{e_y}^2} \quad (3.8)$$

O cálculo do VPC para modelos com resposta binária foi introduzido por Golsdtein et al. (2002) [11], vista a diferença dessa para a regressão linear no cálculo e interpretação de seus resíduos. Para exemplificar, podemos escrever o modelo de regressão logística multinível da seguinte forma:

$$E(y_i) = p_i = \text{logit}^{-1}(\beta X_i + \gamma_{j[i]} W_j + r_{0,j}) \quad (3.9)$$

$$y_i \sim \text{Bernoulli}(p_i) \quad (3.10)$$

$$r_{0,j} \sim N(0, \sigma_{r0}^2) \quad (3.11)$$

Como em regressões lineares buscamos estimar um valor observado, o resíduo e sua variância populacionais são de simples construção com base na diferença entre o valor real e estimado, com $e_y = y_i - \hat{y}_i$ e $e_y \sim N(0, \sigma_e^2)$, e $\text{var}(e_y) = \sum_1^n e_{yi}^2$. Porém, na regressão logística o cálculo da variância dependeria de diretamente de p_i ($\text{var}(e_{yi}) = p_i(1 - p_i)$), não sendo viável a estimação da variância amostral [11].

Os autores apresentam diferentes abordagens para o cálculo de σ_y^2 , utilizaremos aqui o método da variável latente por simplicidade. Neste método assumimos que a variável resposta y possui uma distribuição contínua e a observação de $y = 1$ é obtida quando este valor ultrapassa um determinado limite. A distribuição de probabilidade de y é então dada pela função logística:

$$f(x) = \frac{\exp x}{(1 + \exp x)^2}$$

A partir da qual podemos calcular sua distribuição de probabilidade acumulada como:

$$\int_Y^\infty f(x)dx = \frac{1}{1 + \exp Y}$$

Vale notar que a expressão é semelhante à função logit^{-1} com o preditor Y igual ao agrupamento de estudantes. A variância para a distribuição logística é igual a $\pi^2/3$, assumida como a variância para a estimação do primeiro nível (σ_y^2) [11]. Assim, podemos expressar a proporção de variância atribuída ao agrupamento pela razão:

$$VPC = \frac{\sigma_{r0}^2}{\sigma_{r0}^2 + \frac{\pi^2}{3}} \quad (3.12)$$

Uma vez obtida a métrica, testamos a implementação de um modelo de intercepto aleatório $\beta_{0,j}$ com as variáveis $W_{k,j}$ das escolas como seus preditores, junto às variáveis β_m a nível aluno.

3.4.2 Modelo multinível - Coeficiente aleatório

Conforme apresentado também na seção 2.3, o modelo de coeficiente aleatório nos permite contemplar a variação entre escolas combinada a preditores a nível aluno, variando seus coeficientes. A expressão para o modelo de coeficiente aleatório é dada por:

$$\text{logit}(\hat{p}_i) = \alpha + \beta_{1,j[i]}X_{1,i} + \dots + \beta_{M,j}X_{M,i}, \quad i = 1, \dots, n \quad (3.13)$$

$$\beta_{m,j} = \beta_0 + \sum_{k=1}^K \gamma_k W_{k,j} + r_{0,j}, \quad j = 1, \dots, J \quad (3.14)$$

sendo $\beta_{m,j}$ o coeficiente da variável a nível aluno X_m para a escola j .

Rodamos esse modelo combinando efeitos de variáveis de escola com aluno, como por exemplo, idade da(o) estudante i e nível socioeconômico da escola $j[i]$. A escolha das variáveis a serem combinadas se deu a partir das variáveis de maior efeito identificadas no modelo do intercepto aleatório, afim de explorar ainda mais possíveis relações.

3.5 Random Forest

Modelos como árvores de decisão são uma opção para a modelagem de relações não lineares mais complexas, que não costumam ser capturadas numa regressão. As árvores de decisão são construídas particionando as observações de forma a minimizar a variância dentro dos grupos separados. Cada regra que define uma partição (nó) é criada escolhendo a variável explicativa que melhor separa os dados. Esse processo é realizado novamente para cada partição usando as variáveis restantes até que não haja uma melhor separação do dado - ou não tenham mais variáveis a serem avaliadas.

O modelo de *random forest* é composto por n possíveis árvores de decisão, construídas a partir de diferentes amostras com reposição dos dados originais cujos nós também são escolhidos a partir de amostras aleatórias das variáveis explicativas, e classifica uma observação dada a média dos resultados para cada árvore (ou votação, no caso binário) [13].

Alguns estudos recentes de Alhem et al (2010) [12] introduzem a utilização de modelos de *random forest* para dados hierárquicos através da abordagem multinível. A estrutura do modelo é semelhante à original, acrescentando preditores lineares para cada grupo observado, como mostra a equação 3.15.

$$\begin{cases} \eta_i = \text{logit}(\hat{p}_i) = f(x_i) + b_j W_{j[i]} \\ \hat{y}_{li} = \eta_i + r_i \\ r_i = \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)} \end{cases} \quad (3.15)$$

sendo $f(x_i)$ a função definida através das subdivisões da árvore, na qual o valor para x_i é dado pela folha (último nó) que contempla suas características, $b_j W_{j[i]}$ é o preditor linear referente ao grupo $j[i]$ e y_{li} a linearização de y_i pela expansão de Taylor de primeira ordem. O modelo estima o valor de y_{li} através do algoritmo de maximização de expectativa (EM) para cálculo da estimação de máxima verossimilhança com pesos $w_i = p_i(1 - p_i)$.

Capítulo 4

Resultados

4.1 Análise exploratória

As análises das variáveis explicativas foram realizadas no universo de 146.214 estudantes em 837 escolas estaduais de ensino médio no Rio de Janeiro em 2016, sendo 31,7% evadidas(os) em 2017. A distribuição das variáveis a nível aluno, selecionadas conforme descrito na seção 3.2, são apresentadas na tabela abaixo.

	Absoluto		Percentual	
	Não evadido	Evadido	Não evadido (%)	Evadido (%)
Homem	49389	25461	65.98	34.02
Mulher	50506	20858	70.77	29.23
Amarela	162	90	64.29	35.71
Branca	18402	7863	70.06	29.94
Indígena	50	40	55.56	44.44
Não declarada	46650	21230	68.72	31.28
Parda	26914	13163	67.16	32.84
Preta	7717	3933	66.24	33.76
Utiliza transporte público	72462	29347	71.17	28.83
Não utiliza	27433	16972	61.78	38.22
Abaixo de 17 anos	80631	21393	79.03	20.97
Acima de 17 anos	19264	24926	43.59	56.41
Estuda no mesmo município	65537	30376	68.33	31.67
Estuda num município diferente	34358	15943	68.30	31.70
Rio de Janeiro	35359	17773	66.55	33.45
Região metropolitana	36941	15618	70.28	29.72
Outro município	27595	12928	68.10	31.90

Tabela 4.1: Distribuição das variáveis a nível aluno em função da indicadora de evasão

Podemos notar que a distorção idade-série se destaca entre as variáveis a nível aluno

do Censo avaliadas - basta observarmos o percentual de evadidos. Dentre estudantes com distorção idade-série, mais de 56% evade no ano seguinte, comparado a apenas 20% daquelas(es) que estão na idade adequada. Essa diferença está de acordo com os fatores levantados pela literatura (Tabela 2.1), pois uma idade mais avançada pode alertar tanto para dificuldade de aprendizagem, materializada na reprovação, quanto desinteresse e/ou empecilhos externos que causam o abandono e, posteriormente, a evasão.

A distribuição da idade de alunas(os) evadidos e não evadidos é mostrada em detalhe na Figura 4.1. Vemos nitidamente que se concentram em faixas distintas de idade: 25%¹ das(os) estudantes evadidos têm até 16 anos, o que se inverte para estudantes não evadidas(os) - somente 25%² dessas(es) têm mais de 16 anos.

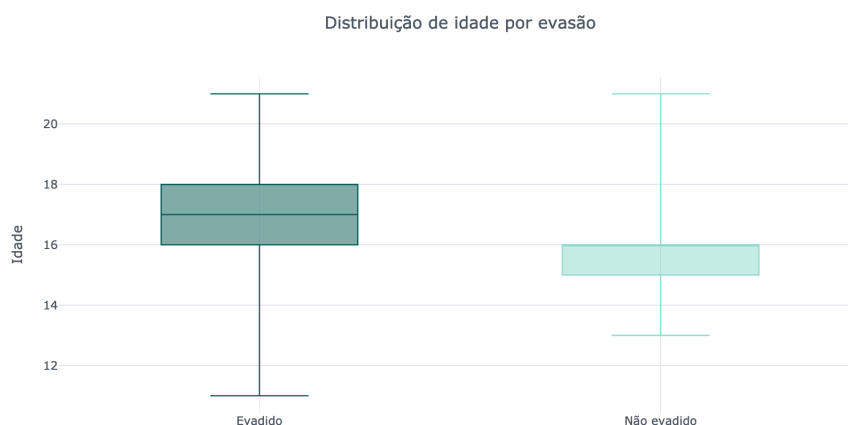


Figura 4.1: Distribuição da idade em função da indicadora de evasão

Outro fator que se destaca na análise dos dados é a utilização do transporte público: 38% de estudantes que não se utilizam do mesmo evadem, 10% a mais dos que o utilizam. Esse fator pode ser observado na literatura através do acesso limitado a infraestrutura, nesse caso especificamente em relação ao transporte (Tabela 2.1). Além disso, o benefício do transporte público gratuito, como o RioCard, serve como um incentivo para a permanência na escola.

Dentre as demais variáveis, nenhuma consegue isoladamente caracterizar o fenômeno da evasão escolar, variando poucos pontos percentuais entre evadidas(os) e não evadidas(os) - por exemplo, o percentual de evadidas(os) que mora no mesmo município da escola não difere daqueles que moram num município diferente, ambos com 31,7%. A diferença entre homens e mulheres observada (cerca de 5% a mais de homens evadem) também não nos fornece muita informação para tratarmos do tema de gravidez na adolescência, dado que o percentual feminino de evasão é menor e não sabemos se, dentre essas, a maioria esteve grávida. Porém, o que pode ocorrer também é um afastamento temporário e posterior retorno das jovens que engravidam durante o período - embora, segundo dados da Pnad de 2014, isso ocorra em apenas 2% país³.

Fatores socioeconômicos e violência não foram possíveis de serem mapeados a nível aluno, por isso trazemos os mesmos a seguir a nível escola. Quanto ao número de estudantes

¹Abaixo do 1º quartil, segmento inferior do retângulo no *boxplot*.

²Acima do 3º quartil, segmento superior do retângulo no *boxplot*.

³Disponível em: <https://www.institutounibanco.org.br/aprendizagem-em-foco/5/>

por escola, vemos na figura 4.3 que 75% das escolas têm até 200 estudantes e a quantidade de escolas decresce exponencialmente com o aumento do número de estudantes - isso nos indica que a distribuição poderia ser aproximada por uma Poisson com parâmetro λ pequeno (a média é de 174 estudantes). A distribuição da taxa de evasão por escola também é apresentada na figura e, como esperado, segue uma distribuição normal com média entorno de 30% e desvio padrão de 11% - isso indica que o fenômeno de evasão está bem distribuído entre as escolas. Apenas 19 escolas têm percentual acima de 50% - todas com favelas e mais de 70% com tiroteios no raio de 1km, e 75% das escolas têm até 36% de evasão.

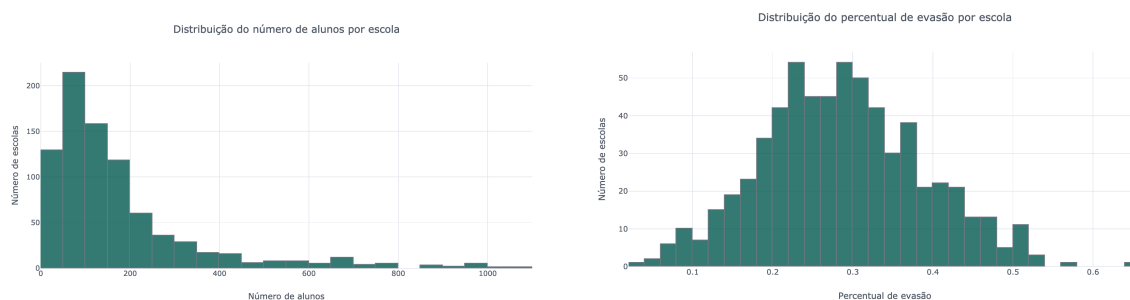


Figura 4.2: Distribuição do número de estudantes e taxa de evasão por escola

A distribuição das variáveis a nível escola (INSE, computadores para estudantes e funcionários por escola) é apresentada na Figura 4.3 em função do percentual de evasão. Vemos que não há uma clara correlação linear do fenômeno com esses fatores. Uma possível explicação para a ausência de relação observada do nível socioeconômico com a evasão, que não condiz com a literatura no fator pobreza (Tabela 2.1), é o cálculo desse indicador socioeconômico, que normaliza os valores fazendo com que estes estejam concentrados entorno da média e acabam não traduzindo a realidade socioeconômica de muitos estudantes.



Figura 4.3: Distribuição das variáveis numéricas

Na tabela 4.2 temos as variáveis de escola (indicadoras) em função da taxa de evasão. Podemos notar que características de infraestrutura das escolas, diferente do que é evidenciado na literatura (Tabela 2.1), não distinguem estudantes evadidas(os) e não evadidos, todas com um percentual e desvio padrão muito próximos. Logo, optamos por incluir somente as variáveis relacionadas ao ambiente da escola (quadra, laboratório, auditório, etc), e desconsiderar variáveis de equipamentos por não serem destacadas na literatura.

	Número de escolas	Taxa de evasão (Desvio padrão (%))
Abastecimento de água (Cacimba/Cisterna)	244	29.65 (9.11)
Outros	593	29.74 (11.14)
Possui laboratório de ciências	621	30.28 (10.57)
Não possui	216	28.1 (10.48)
Possui quadra descoberta	588	30.23 (10.59)
Não possui	249	28.51 (10.48)
Possui banheiro fora do prédio	643	30.3 (10.42)
Não possui	194	27.79 (10.91)
Possui auditório	664	30.3 (10.6)
Não possui	173	27.48 (10.22)
Possui alojamento para aluna(o)	457	30.16 (9.95)
Não possui	380	29.18 (11.29)
Possui videocassete	238	29.28 (10.72)
Não possui	599	29.89 (10.53)
Possui antena parabólica	404	28.95 (10.25)
Não possui	433	30.43 (10.85)
Possui retroprojektor	439	29.65 (10.14)
Não possui	398	29.79 (11.07)
Possui impressora	648	29.3 (10.36)
Não possui	189	31.13 (11.22)
Possui equipamento de fax	302	28.17 (10.2)
Não possui	535	30.59 (10.7)
Oferta de EJA	393	31.89 (10.2)
Não possui	444	27.79 (10.55)
Oferta de E.F.	590	28.51 (10.23)
Não possui	247	32.6 (10.87)
Comunidade num raio de 1KM	480	31.35 (11.11)
Sem comunidade próxima	357	27.52 (9.4)
Comunidade num raio de 5KM	642	30.95 (10.61)
Sem comunidade próxima	195	25.64 (9.42)
Tiroteio num raio de 1KM	328	30.84 (11.11)
Sem tiroteio registrado	509	28.99 (10.17)
Tiroteio num raio de 5KM	456	30.54 (10.75)
Sem tiroteio registrado	381	28.73 (10.3)
Rio de Janeiro	93	34.71 (14.02)
Região metropolitana	405	29.68 (9.44)
Outro município	339	28.39 (10.41)

Tabela 4.2: Distribuição das variáveis a nível aluno em função da indicadora de evasão

4.2 Efeito escola

O primeiro modelo avaliado foi o modelo nulo, adaptado da equação 4.2 para o modelo logístico como segue na equação abaixo. Neste modelo, temos somente um intercepto $\alpha_{0,j}$ composto de um fator fixo α_0 e um resíduo que varia por escola $r_{0,j}$.

$$\text{logit}(\hat{p}_i) = \alpha_{0,j[i]} \quad (4.1)$$

$$\alpha_{0,j} = \alpha_0 + r_{0,j}, \quad j = 1, \dots, J \quad (4.2)$$

A estimativa do intercepto fixo α_0 calculada a partir dos dados de treino segue na Tabela 4.3, significativa a menos de 1% ($p\text{-valor} = 6.75e - 07$).

	Estimativa	Erro padrão
α_0	-0.8878	0.0206

Tabela 4.3: Estimativa do coeficiente de primeiro nível do modelo nulo

Esse valor nos indica que a probabilidade média de evasão é dada por $\exp -0.888 \approx 0.41$. Os resíduos do segundo nível $r_{0,j}$ e seus intervalos de confiança são apresentados na Figura 4.4. Podemos notar que quase todos os intervalos incluem o zero, ou seja, a estimativa da evasão para essas escolas não difere significativamente da média - dada por α_0 . Esse é um resultado razoável, dado que a distribuição da taxa de evasão para as escolas do nosso universo 4.3 segue uma distribuição normal centrada em 30%.

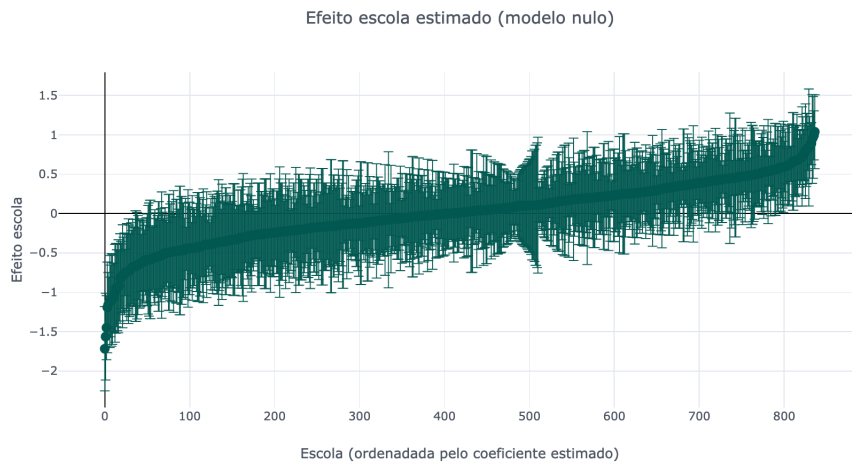


Figura 4.4: Efeito escola estimado no modelo nulo

A estimativa da variância residual das escolas ($\sigma_{r_0}^2$) é de 0.2, o que nos fornece $VPC \approx 0.06$. Um valor próximo de zero indica que o agrupamento em si não traz muita informação - enquanto para um $VPC \approx 1$ observações de um mesmo grupo são praticamente idênticas [10]. Para fins de comparação, no estudo realizado por Shirasu e Arraes [16] com escolas do Ceará, a variância residual por escola correspondia a cerca de 20%, indicando que fatores a nível escola tinham maior efeito na evasão.

Como o percentual de evasão das escolas não varia muito entre escolas, é natural que o modelo nulo não seja um bom preditor desse fenômeno, pois a estimativa gerada é constante para alunas(os) de uma mesma escola. A figura 4.5 mostra o cálculo da precisão e *recall* para diferentes limites δ escolhidos para classificação (eixo x), segundo a fórmula abaixo.

$$\hat{y}_i^{(\delta)} = \begin{cases} 1, & \text{se } \hat{p}_i \geq \delta \\ 0, & \text{c.c.} \end{cases} \quad (4.3)$$

A precisão é dada pela razão de alunas(os) classificados corretamente como evadidas(os) pelo total de classificados nessa situação⁴, enquanto o *recall* mensura, dentre estudantes evadidas(os), quantas(os) são classificadas(os) corretamente⁵. Para comparação, apresentamos também na Figura 4.5 a classificação dada uma estimativa aleatória de p_i - uma performance semelhante a essa seria o pior cenário para o modelo, pois não estaríamos adicionando nenhuma informação relevante para a previsão com a inserção das variáveis escolhidas. Como podemos observar, o modelo nulo não apresenta um ganho elevado em relação ao aleatório: de 20% das(os) estudantes estimadas(os) com maior probabilidade⁶, temos uma precisão de apenas 40% e *recall* de aproximadamente 25% - apenas 5% a mais que o aleatório.

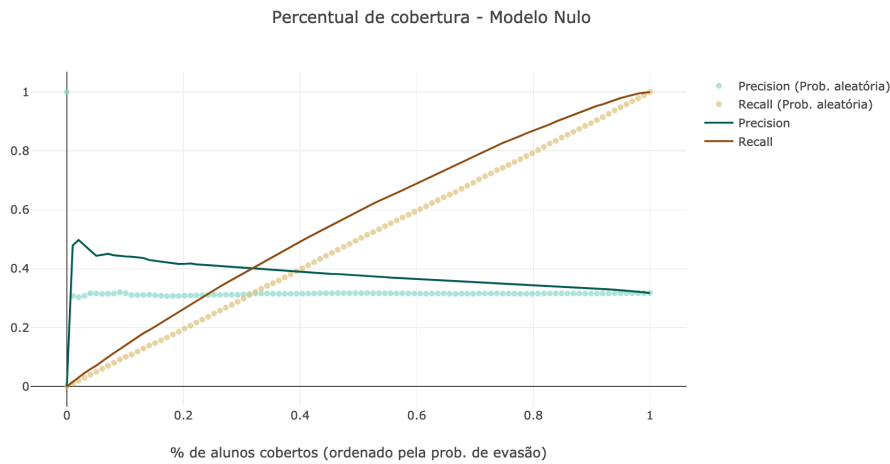


Figura 4.5: Percentual de cobertura do modelo nulo

Com a adição das variáveis de escola selecionadas, o resultado das predições não apresenta diferença. Esse também é um resultado razoável, dado que os percentuais de evasão para cada categoria são muito próximos (conforme mostrado Tabela 4.2). Os coeficientes estimados para os preditores seguem na Tabela 4.4 - significantes a menos de 5%, com exceção da indicadora da escola na cidade Rio de Janeiro. Podemos observar que a proximidade de comunidade num raio de 5km tem efeito positivo na evasão e, não muito significativa mas a existência de laboratório de ciências também se apresenta positiva. Por outro lado, o nível socioeconômico (INSE), quantidade de computadores na escola e a localização na região metropolitana⁷ se mostram fatores com efeito negativo. Para uma estimativa da magnitude desses efeitos, tomemos como exemplo o nível socioeconômico: o coeficiente $\gamma_6 = -0.125$

⁴Verdadeiro positivos / (Falso positivos + Verdadeiro positivos)

⁵Verdadeiro positivos / (Falso negativos + Verdadeiro positivos)

⁶Indicado pelo ponto 0.2 no eixo x

⁷As variáveis numéricas foram normalizadas, com média igual a 0 e desvio padrão igual a 1

Coeficiente	Variável	Estimativa	Erro padrão
$\alpha_{0,j}$	Intercepto fixo	-1.1785	0.0487
γ_1	L2_IN_MUNICIPIO_RIO	0.1039	0.0607
γ_2	L2_IN_MUNICIPIO_MET	-0.1247	0.0412
γ_3	L2_IN_LABORATORIO_CIENCIAS	0.0887	0.0393
γ_4	L2_NU_FUNCIONARIOS_NORM	-0.0379	0.0187
γ_5	L2_NU_COMP_ALUNO_NORM	-0.1020	0.0181
γ_6	L2_NU_INSE_VALOR_NORM	-0.1249	0.0186
γ_7	L2_IN_COMUNIDADE_5KM	0.3581	0.0508

Tabela 4.4: Estimativa dos coeficientes do modelo com variáveis de escola

estima que a variação de uma unidade de desvio padrão (≈ 2.53 unidades no valor do INSE) diminui a probabilidade de evasão em $\approx 3\%$ ⁸. De forma semelhante, a proximidade da escola a alguma comunidade num raio de 5km aumenta a probabilidade de evasão em aproximadamente 8.75%.

4.3 Modelo base: Idade

Dada a relevância do fator idade citada na literatura e corroborado para o nosso contexto através da análise exploratória dos dados (conforme a Tabela 4.1), vamos assumir como modelo base para comparação das previsões a regressão estimada somente com essa variável. Logo, buscamos mensurar o poder preditivo da idade no contexto apresentado e entender se a adição variáveis consegue melhorar esses resultados.

Na Tabela 4.5 constam os coeficientes estimado no modelo. A idade apresenta-se como um forte fator positivo para a evasão: estima-se que a cada ano mais velho a probabilidade de evasão pode aumentar em até 23%⁹.

Coeficiente	Variável	Estimativa	Erro padrão
α_0	Intercepto fixo	-0.8447	0.0072
β_1	L1_NU_IDADE_REFERENCIA_NORM	0.9220	0.0092

Tabela 4.5: Estimativa dos coeficientes do modelo base

A Figura 4.6 mostra a estimativa das probabilidades para estudantes evadidas(os) e não evadidas(os) na base de teste. Como existem poucos valores de idade diferentes observados na base, as possibilidades de valores para estimação da probabilidade são poucas. Entretanto, o modelo consegue atribuir probabilidades mais baixas para estudantes não evadidas(os) com maior frequência, e vemos que a proporção de estudantes evadidas(os) cresce conforme o aumento da probabilidade estimada.

Essa distribuição nos indica que provavelmente o modelo consegue classificar corretamente estudantes evadidas(os) de forma melhor que o anterior, o que se mostra verdadeiro na Figura 4.7. Dentre 20% das(os) alunas(os) classificadas(os) com maior probabilidade em

⁸Próximo ao ponto $\text{logit}(0.5) = \gamma_6/4$

⁹Desvio padrão ≈ 1.2 , $\beta_1/4 \approx 0.23$

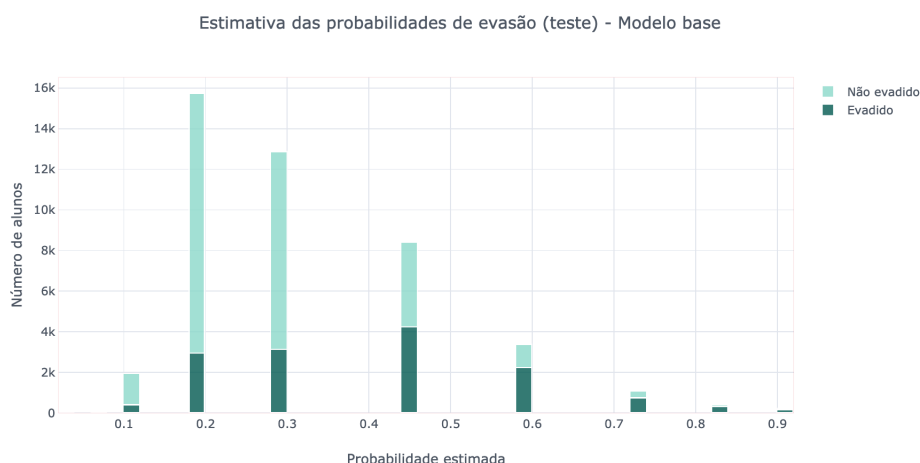


Figura 4.6: Probabilidade estimada pelo modelo base

ambos os modelos, o modelo base apresenta um *recall* de $\approx 38\%$ e precisão de 70% - um ganho de respectivamente 13% e 30% comparado ao modelo nulo. Isso significa que, dentre as(os) 20% classificados com maior risco de evasão, estima-se que 70% efetivamente evada, cobrindo aproximadamente 38% do total de alunas(os) evadidas(os).

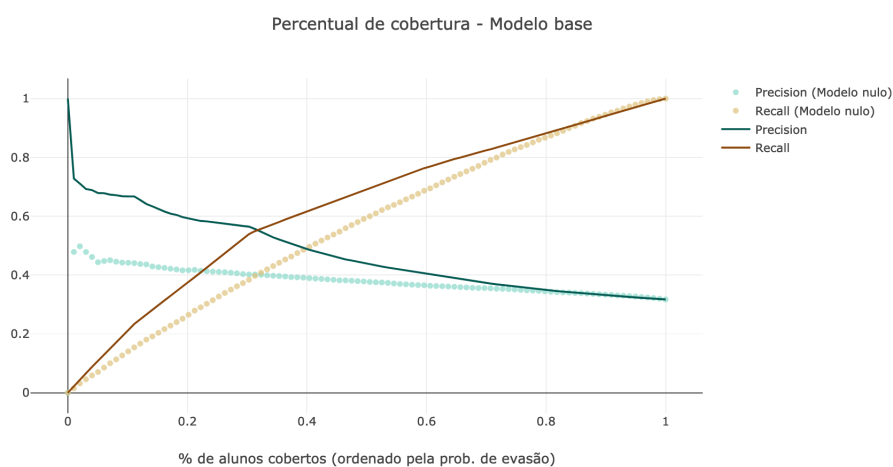


Figura 4.7: Percentual de cobertura do modelo base (idade)

4.4 Intercepto e coeficiente aleatório

O primeiro modelo testado com a combinação das variáveis de estudantes e escolas foi o intercepto aleatório, no qual permitimos que o intercepto varie em função das características das escolas. Pela implementação do pacote *lmer* em *R*, as variáveis de escola são passadas da mesma forma que variáveis de alunas(os) e determinamos o agrupamento em função dos códigos das escolas - como $\gamma_1, \dots, \gamma_K$ são fixos dado j , passá-los como variáveis do primeiro

nível na prática nos fornece um modelo de intercepto aleatório¹⁰.

A Tabela 4.6 mostra os coeficientes estimados para cada variável do modelo. Os primeiros na tabela, α_0, β_k , são referentes aos efeitos fixos estimados (variáveis do nível aluno(o)). Comparando ao modelo base, as variáveis adicionadas têm efeito negativo nas chances de evasão. O efeito estimado da idade se mantém como aproximadamente 23,8%, enquanto a utilização de transporte público apresenta uma estimativa de diminuição de 9% da probabilidade de evasão.

Coeficiente	Variável	Estimativa	Erro padrão
α_0	Intercepto fixo	-1.0644	0.0499
β_1	L1_IN_SEXO (Mulher)	-0.0727	0.0147
β_2	L1_IN_TRANSPORTE_PUBLICO	-0.3612	0.0174
β_3	L1_IN_MUNICIPIO_NASC_RIO	-0.0528	0.0220
β_4	L1_NU_IDADE_REFERENCIA_NORM	0.9348	0.0098
γ_1	L2_IN_MUNICIPIO_RIO	0.3483	0.0546
γ_2	L2_IN_LABORATORIO_Ciencias	0.1047	0.0394
γ_3	L2_NU_FUNCIONARIOS	-0.0319	0.0185
γ_4	L2_NU_COMP_ALUNO	-0.1149	0.0180
γ_5	L2_NU_INSE_VALOR	-0.1001	0.0186
γ_6	L2_IN_COMUNIDADE_5KM	0.2859	0.0451

Tabela 4.6: Estimativa dos coeficientes do modelo com intercepto aleatório

O efeito negativo do sexo feminino também é observado no modelo de intercepto aleatório trabalhado por Shiratu et al. (2016) [16] no Ceará¹¹. A idade mais velha relacionada à evasão, identificada no estudo por uma indicadora de atraso escolar, também tem efeito positivo - além dessa, a indicadora de repetência anterior se destaca com um dos maiores efeitos estimados. O nível socioeconômico, porém, tem efeito positivo, o que pode ocorrer devido ao fato de usarem a classificação dos níveis (1 a 7) ao invés do valor absoluto calculado pelo INEP, o que diminui a diferenciabilidade entre escolas. Embora sejam efeitos opostos, o efeito estimado do INSE no estudo é ainda próximo de zero - apenas 0.016.

Comparando as variáveis de escola (efeitos aleatórios) ao modelo composto somente do agrupamento das escolas (vide Tabela 4.4, obtemos efeitos também semelhantes, o que corrobora para a consistência do modelo. Porém, isso significa que não iremos observar muita diferença com a inserção dessas variáveis, dado que elas não acrescentam poder de previsão ao agrupamento por escola, como constatado na adição das mesmas ao modelo nulo (Seção 4.2)) e apresentado na Figura 4.8. Vemos que o ganho de performance é marginal, tendo uma precisão de poucos pontos percentuais a mais para as(os) alunas(os) classificados dentre os 10% com maior risco.

¹⁰Dado um agrupamento que torna esses valores constantes.

No *lmer*, a expressão se assemelha a: $y \sim x_1 + \dots + x_M + w_1 + \dots + w_K + (1|J[I])$

¹¹Sexo de referência no estudo é o masculino, logo o efeito estimado é positivo

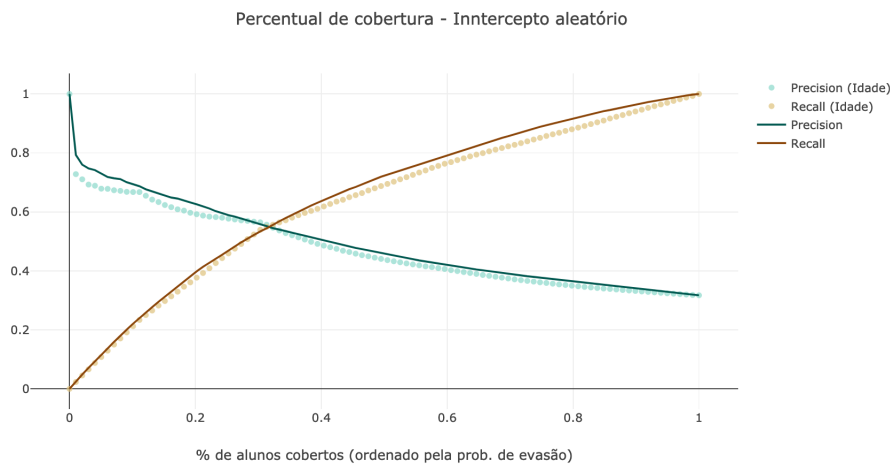


Figura 4.8: Percentual de cobertura do modelo com intercepto aleatório

A distribuição da probabilidade estimada pelo modelo segue na Figura 4.9. Vemos que estudantes não evadidos(os) estão concentrados em probabilidades mais baixas, o que indica que nosso modelo consegue classificar bem essa categoria. Porém, a distribuição de probabilidade estimada para estudantes evadidos(os) ainda não consegue capturar razoavelmente este fenômeno, dado que as probabilidades se concentram também em valores mais baixo.

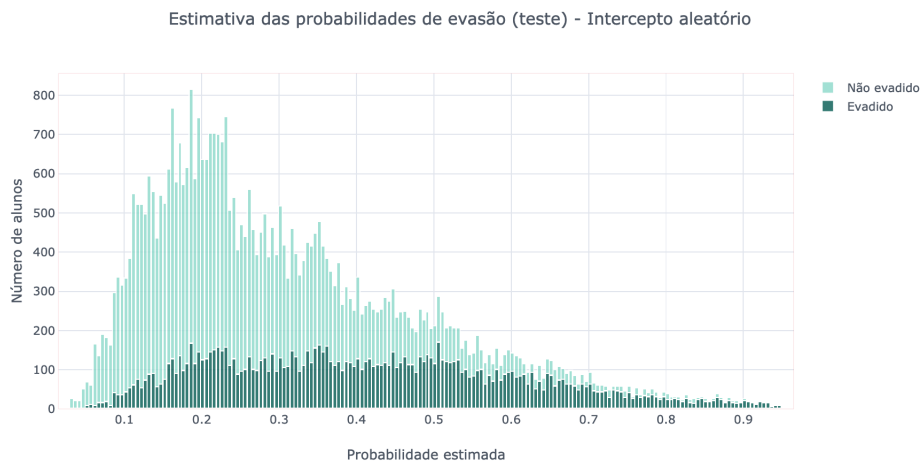


Figura 4.9: Probabilidade estimada pelo intercepto aleatório

A variância residual por escolas é de $\sigma_{r_0}^2 \approx 0.15$, menor que a estimada pelo modelo nulo dada a introdução de características da escola. Como visto na seção 3.4.1, o intercepto aleatório é calculado por $\alpha_{0,j} = \sum_k \gamma_k W_k + r_{0,j}$. A Figura 4.10 apresenta a estimação por escola - não difere muito do coeficiente observado no modelo nulo ¹², dado o que vimos na análise exploratória (vide Tabela 4.2) que o percentual de evasão não difere significativamente por fatores da escola.

A partir desses resultados, selecionamos as variáveis de escola com maiores coeficientes para combinarmos a variáveis de estudantes com a inserção de coeficientes aleatórios, porém

¹²As escalas estão ligeiramente diferentes: o eixo y está no intervalo $[-2, 2]$ e, no modelo nulo, $[-2, 1.5]$

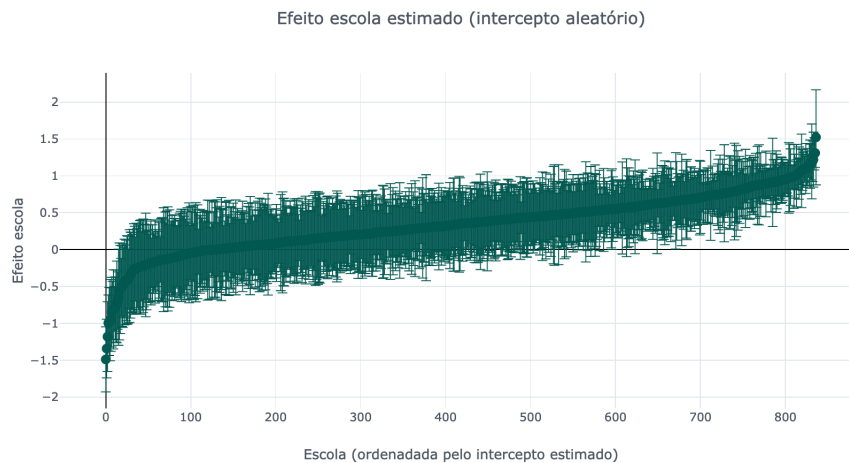


Figura 4.10: Intercepto aleatório estimado por escola

não tivemos sucesso na convergência do modelo. A introdução de relações entre variáveis torna mais complexo o cálculo dos seus coeficientes e pode não ter um número representativo de observações para cada subdivisão - é o caso de termos escolas com poucas alunas(os) e possivelmente nenhuma ou total evasão.

4.5 *Mixed Random Forest*

Por fim, apresentamos os resultados do modelo hierarquizado de *random forest*. As probabilidades estimadas seguem na Figura 4.11. Conseguimos ver claramente um fator separador relevante no modelo que perpassa ambas as classes, e caracteriza bem estudantes não evadidos(as).

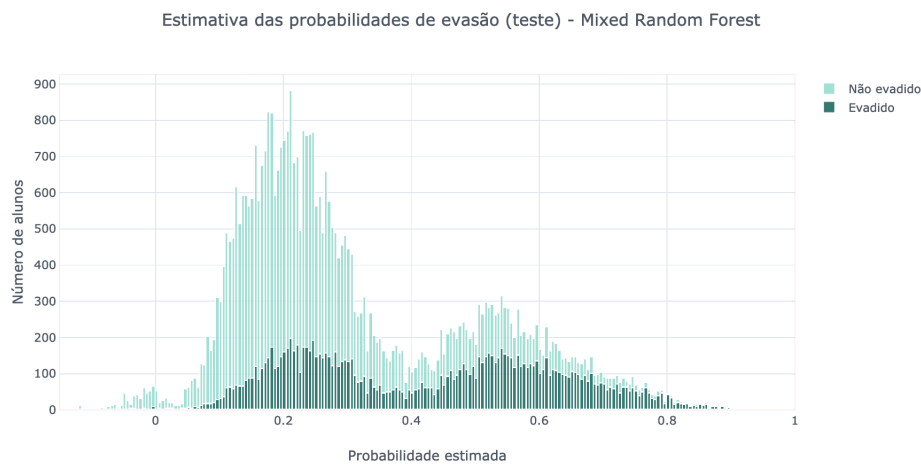


Figura 4.11: Probabilidade estimada pelo *mixed random forest*

A Tabela 4.7 mostra a quantidade de estudantes classificados com probabilidade abaixo e acima de 0.4, separados pelas variáveis a nível aluno e escola. Identificamos, como de se esperar, que a idade é o fator separador primário dos dados, pois é o fator que melhor

caracteriza o fenômeno. Os percentuais das demais variáveis não se mostram muito distintos entre estudantes classificadas(os) com maior ou menor probabilidade.

	p =< 0.4	p > 0.4	p =< 0.4 (%)	p > 0.4 (%)
Homem	14710	7777	47.69	59.73
Mulher	16135	5243	52.31	40.27
Utiliza transporte público	22658	7854	73.46	60.32
Não utiliza	8187	5166	26.54	39.68
Abaixo de 17 anos	30220	286	97.97	2.20
Acima de 17 anos	625	12734	2.03	97.80
Estuda no mesmo município	20371	8382	66.04	64.38
Estuda num município diferente	10474	4638	33.96	35.62
Possui laboratório de ciências	24275	10325	78.70	79.30
Não possui	6570	2695	21.30	20.70
Comunidade num raio de 5KM	27543	11806	89.29	90.68
Sem comunidade próxima	3302	1214	10.71	9.32
Rio de Janeiro	9132	3568	29.61	27.40
Região metropolitana	13702	6190	44.42	47.54
Outro município	8011	3262	25.97	25.05

Tabela 4.7: Separação das observações no *mixed random forest*

A performance desse comparado ao modelo base tem um resultado semelhante ao encontrado para o modelo com intercepto aleatório (vide Figura 4.8. Como mostra a Figura 4.12, o ganho não é muito significativo. O distanciamento observado a partir de 0.4 condiz com a distribuição mostrada na figura anterior, pois existe uma maior concentração de estudantes classificadas(os) como evadidas(os) nessa faixa - isso é verdade também quando comparado ao intercepto aleatório.

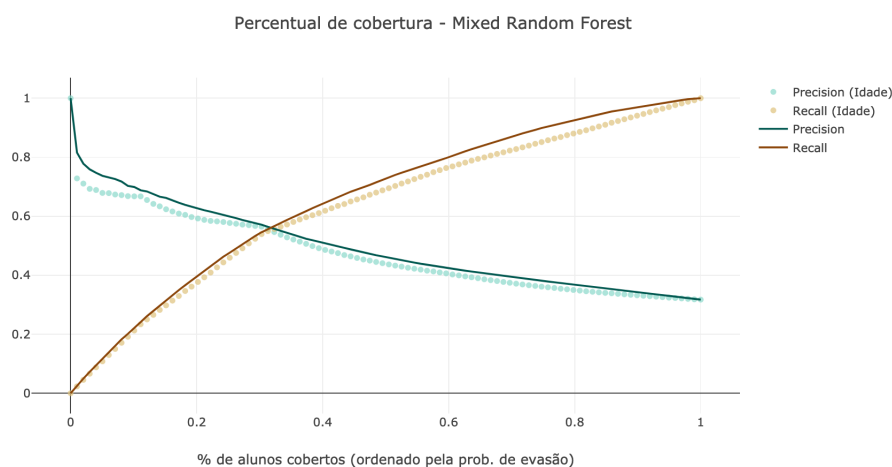


Figura 4.12: Percentual de cobertura do modelo de *mixed random forest*

Para fins de comparação entre todos os modelos, utilizamos a curva ROC (*Receiver Operating Characteristic*), construída como taxa de verdadeiros positivos (sensibilidade) versus a taxa de falsos positivos (especificidade) para vários limites ordenados de classificação escolhidos em $\delta \in [0, 1]$ ¹³ [8]. Quando o modelo classifica perfeitamente os dados, a curva ROC passa pelo ponto $(0, 1)$ com $\delta = \delta^*$ (separador completo dos dados), como demonstrado abaixo.

$$FN_{\delta^*} = 0 \rightarrow \begin{cases} \frac{VP_{\delta^*}}{VP_{\delta^*} + FN_{\delta^*}} = 1 \\ 1 - \frac{VN_{\delta^*}}{VN_{\delta^*} + FN_{\delta^*}} = 0 \end{cases} \quad (4.4)$$

Em contrapartida, a diagonal ($y = x$) representa a curva ROC de um modelo aleatório (pior cenário), no qual a proporção de estudantes evadidas(os) corretamente classificadas(os) é igual à proporção de estudantes não evadidas(os) incorretamente classificadas (como evadidas). Logo, para um dado modelo, quanto mais a curva se afasta da diagonal e se aproxima de $(0, 1)$, melhor a classificação do mesmo. A área sob da curva ROC (AUC) é um valor que sintetiza a qualidade do modelo: a área é equivalente à probabilidade do modelo classificar uma aluna(o) que evadiu com maior probabilidade de evasão do que uma aluna(o) não evadiu [8].

A curva ROC de cada modelo trabalhado é apresentada na Figura 4.13. Podemos constatar, como visto nas seções anteriores, que o ganho de performance dos modelos em relação ao modelo base (idade) é mínimo. Isso se mostra também no cálculo da área sob a curva (AUC), cujo valor máximo dentre os modelos é obtido pelo *mixed random effect*, com $AUC = 0.74$, porém a diferença para o intercepto aleatório ($AUC = 0.73$) e o modelo base ($AUC = 0.7$) é praticamente irrelevante.

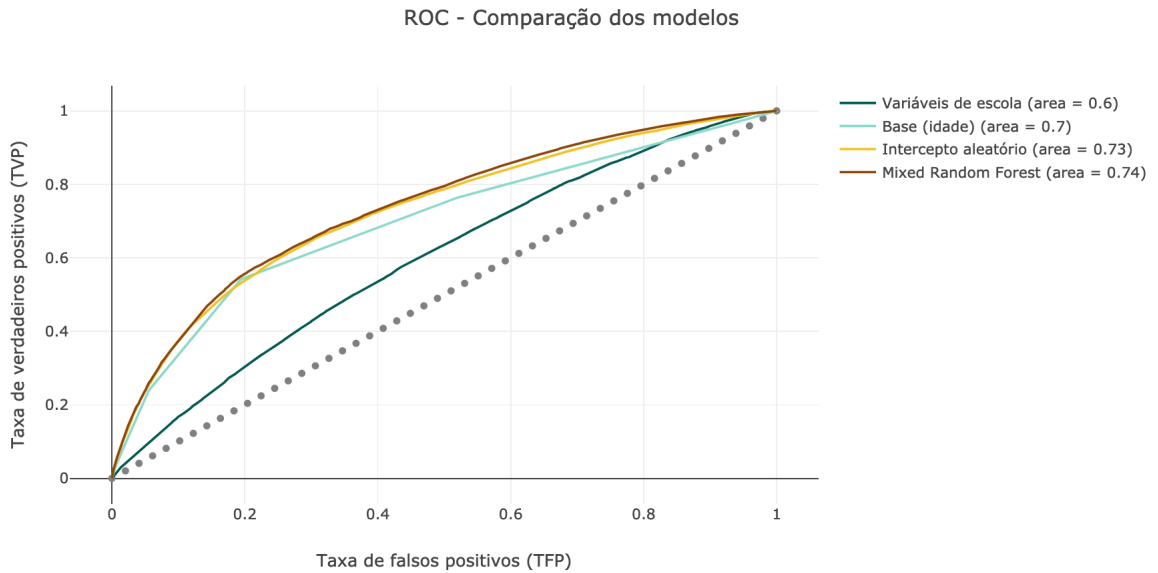


Figura 4.13: Curvas ROC dos modelos trabalhados

¹³Abordagem da variável latente: o limite δ define $y_i = \{1, \text{se } \hat{p}_i \geq \delta, 0 \text{ c.c.}\}$

Capítulo 5

Conclusões

No presente trabalho buscamos prever o fenômeno da evasão escolar no 1º ano do ensino médio estadual no Rio de Janeiro através dos dados do Censo Escolar. A análise exploratória dos dados revelou uma forte relação da idade da(o) aluna(o) com a evasão, que utilizamos como modelo base, e certa homogeneidade das variáveis de escola para alunas(os) evadidas(os) e não evadidas(os), o que se comprovou nos modelos testados. O baixo valor obtido do coeficiente de variância particionada (VPC) das escolas também nos indica que o fenômeno da evasão está pouco relacionado ao ambiente escolar em si, mas sim a característica para além dessas - seja a nível aluno ou turma.

As performances dos modelos de *mixed random forest* e intercepto aleatório não diferiram significativamente, e tiveram um ganho marginal em relação ao modelo base. Entretanto, a vantagem de se utilizar modelos mais complexos como esses está na incorporação de outros fatores relacionados à evasão que podem ser relevantes para decisões de políticas públicas, ou que podem melhor explorados a nível local.

A principal limitação e possível melhoria desse trabalho está nas informações utilizadas para a construção dos modelos. O INEP divulga uma série de dados e indicadores educacionais de diferentes níveis, alguns que poderiam ser adicionados a nível escola e turma, como nível de formação de professores, mas a nível aluno ainda é escasso. Os dados de desempenho e indicadores socioeconômicos de alunas(os), medidos e coletados pelo SAEB, não são comparáveis aos dados do Censo por não terem um identificador comum, e, a nível escola, têm baixo de preenchimento (menos de 75% das escolas tinham dados de desempenho). Além disso, trabalhar com dados históricos de fatores como repetência e mudanças estruturais podem ajudar a elucidar o fenômeno e avaliar riscos de evasão a longo prazo.

A tentativa de prever a evasão escolar deve ser considerado como um passo importante para não só combater a esse problema, mas também como um auxílio para o diagnóstico de desigualdades educacionais e mensuração de possíveis efeitos de políticas públicas. Com aperfeiçoamentos e testes de diferentes modelos, a previsão atrelada a ferramentas como o Busca Ativa Escolar poderão aumentar ainda mais o impacto dessas iniciativas, permitindo a indentificação e tratamento de alunas(os) com maior risco e possibilitando o sucesso de suas trajetórias na escola.

Apêndice A

Pré-processamento das variáveis de matrícula do Censo Escolar

Tabela A.1: Pré-processamento das variáveis de alunas(os)

	Variável	Tipo	Seleção
1	NU_IDADE_REFERENCIA ¹	Numérica	Variáveis utilizadas (9)
2	TP_SEXO	Catégorica	
3	TP_COR_RACA	Catégorica	
4	CO_MUNICIPIO_NASC ²	Catégorica	
5	IN_TRANSPORTE_PUBLICO	Indicadora	
6	ID_TURMA	Identificadora	
7	CO_ENTIDADE	Catégorica	
8	ID	Identificadora	
9	IN_EVASAO	Indicadora	
10	NU_DIA	Numérica	Informação duplicada (4)
11	NU_MES	Numérica	
12	NU_ANO	Numérica	
13	NU_IDADE	Numérica	
14	CO_MUNICIPIO	Catégorica	Informação de escolas (4)
15	CO_MESORREGIAO	Catégorica	
16	CO_MICRORREGIAO	Catégorica	
17	CO_DISTRITO	Catégorica	
18	CO_UF_END	Catégorica	Baixo preenchimento - < 30% (35)
19	CO_MUNICIPIO_END	Catégorica	
20	IN_CEGUEIRA	Indicadora	
21	IN_BAIXA_VISAO	Indicadora	
22	IN_SURDEZ	Indicadora	
23	IN_DEF_AUDITIVA	Indicadora	
24	IN_SURDOCEGUEIRA	Indicadora	
25	IN_DEF_FISICA	Indicadora	
26	IN_DEF_INTELECTUAL	Indicadora	
27	IN_DEF_MULTIPLA	Indicadora	
28	IN_AUTISMO	Indicadora	
29	IN_SINDROME_ASPIRGER	Indicadora	
30	IN_SINDROME_RETT	Indicadora	
31	IN_TRANSTORNO_DI	Indicadora	
32	IN_SUPERDOTACAO	Indicadora	
33	IN_RECURSO_LEDOR	Indicadora	
34	IN_RECURSO_TRANSCRICAO	Indicadora	
35	IN_RECURSO_INTERPRETE	Indicadora	
36	IN_RECURSO_LIBRAS	Indicadora	
37	IN_RECURSO_LABIAL	Indicadora	
38	IN_RECURSO_BRILLE	Indicadora	

¹Remoção da cauda (< 99-quantil)

²Categorização binária (Rio / Região Metropolitana / Outros)

36 APÊNDICE A. PRÉ-PROCESSAMENTO DAS VARIÁVEIS DE MATRÍCULA DO CENSO ESCOLAR

39	IN_RECURSO_AMPLIADA_16	Indicadora	
40	IN_RECURSO_AMPLIADA_20	Indicadora	
41	IN_RECURSO_AMPLIADA_24	Indicadora	
42	IN_RECURSO_NENHUM	Indicadora	
43	TP_INGRESSO_FEDERAIS	Catégorica	
44	CO_CURSO_EDUC_PROFISSIONAL	Catégorica	
45	TP_CATEGORIA_ESCOLA_PRIVADA	Catégorica	
46	IN_CONVENIADA_PP	Indicadora	
47	TP_CONVENIO_PODER_PUBLICO	Catégorica	
48	IN_MANT_ESCOLA_PRIVADA_EMP	Indicadora	
49	IN_MANT_ESCOLA_PRIVADA_ONG	Indicadora	
50	IN_MANT_ESCOLA_PRIVADA_SIND	Indicadora	
51	IN_MANT_ESCOLA_PRIVADA_SIST_S	Indicadora	
52	IN_MANT_ESCOLA_PRIVADA_S_FINS	Indicadora	
<hr/>			
53	NU_DURACAO_TURMA	Numérica	Baixa variação - > 90% concentrado (40)
54	NU_ANO_CENSO	Numérica	
55	NU_DUR_ATIV_COMP_MESMA_REDE	Numérica	
56	NU_DUR_ATIV_COMP_OUTRAS_REDES	Numérica	
57	NU_DUR_AEE_MESMA_REDE	Numérica	
58	NU_DUR_AEE_OUTRAS_REDES	Numérica	
59	NU_DIAS_ATIVIDADE	Numérica	
60	TP_NACIONALIDADE	Catégorica	
61	CO_PAIS_ORIGEM	Catégorica	
62	CO_UF_NASC	Catégorica	
63	TP_ZONA_RESIDENCIAL	Catégorica	
64	TP_OUTRO_LOCAL_AULA	Catégorica	
65	TP_RESPONSAVEL_TRANSPORTE	Catégorica	
66	IN_TRANSP_VANS_KOMBI	Indicadora	
67	IN_TRANSP_MICRO_ONIBUS	Indicadora	
68	IN_TRANSP_ONIBUS	Indicadora	
69	IN_TRANSP_BICICLETA	Indicadora	
70	IN_TRANSP_TR_ANIMAL	Indicadora	
71	IN_TRANSP_OUTRO_VEICULO	Indicadora	
72	IN_TRANSP_EMBAR_ATE5	Indicadora	
73	IN_TRANSP_EMBAR_5A15	Indicadora	
74	IN_TRANSP_EMBAR_15A35	Indicadora	
75	IN_TRANSP_EMBAR_35	Indicadora	
76	IN_TRANSP_TREM_METRO	Indicadora	
77	IN_NECESIDADE_ESPECIAL	Indicadora	
78	TP_MEDIACAO_DIDATICO_PEDAGO	Catégorica	
79	IN_ESPECIAL_EXCLUSIVA	Indicadora	
80	IN_REGULAR	Indicadora	
81	IN_EJA	Indicadora	
82	IN_PROFISSIONALIZANTE	Indicadora	
83	TP_ETAPA_ENSINO	Catégorica	
84	TP_UNIFICADA	Catégorica	
85	TP_TIPO_TURMA	Catégorica	
86	CO_REGIAO	Catégorica	
87	CO_UF	Catégorica	
88	TP_DEPENDENCIA	Catégorica	
89	TP_LOCALIZACAO	Catégorica	
90	TP_REGULAMENTACAO	Catégorica	
91	TP_LOCALIZACAO_DIFERENCIADA	Catégorica	
92	IN_EDUCACAO_INDIGENA	Indicadora	

Apêndice B

Pré-processamento das variáveis de escola do Censo Escolar

Tabela B.1: Pré-processamento das variáveis de escola

	Variável	Tipo	Seleção
CO_ENTIDADE	Catégorica	Variáveis utilizadas (29)	
CO_ORGAO_REGIONAL	Catégorica		
IN_AGUA_CACIMBA	Indicadora		
IN_LABORATORIO_Ciencias	Indicadora		
IN_QUADRA_ESPORTES_DESCOBERTA	Indicadora		
IN_BANHEIRO_FORA_PREDIO	Indicadora		
IN_AUDITORIO	Indicadora		
IN_ALOJAM_ALUNO	Indicadora		
IN_EQUIP_VIDEOCASSETE	Indicadora		
IN_EQUIP_PARABOLICA	Indicadora		
IN_EQUIP_RETROPROJETOR	Indicadora		
IN_EQUIP_IMPRESSORA	Indicadora		
IN_EQUIP_FAX	Indicadora		
NU_EQUIP_TV	Númerica		
NU_EQUIP_DVD	Númerica		
NU_EQUIP_COPIADORA	Númerica		
NU_EQUIP_SOM	Númerica		
NU_EQUIP_MULTIMIDIA	Númerica		
NU_EQUIP_FOTO	Númerica		
NU_COMP_ALUNO	Númerica		
NU_FUNCIONARIOS	Númerica		
IN_EJA	Indicadora		
INSE_VALOR_ABSOLUTO	Indicadora		
IN_COMUNIDADE_1KM	Indicadora		
IN_COMUNIDADE_5KM	Indicadora		
IN_TIROLEIO_1KM	Indicadora		
IN_TIROLEIO_5KM	Indicadora		
NU_APROV_2015_9	Númerica		
NU_IDEB_2015	Númerica		
NO_ENTIDADE	Nominal	Informação duplicada (10)	
CO_MESORREGIAO	Catégorica		
CO_MICRORREGIAO	Catégorica		
CO_MUNICIPIO	Catégorica		
CO_DISTrito	Catégorica		
NU_EQUIP_VIDEOCASSETE	Númerica		
NU_EQUIP_PARABOLICA	Númerica		
NU_EQUIP_RETROPROJETOR	Númerica		
NU_EQUIP_IMPRESSORA	Númerica		
NU_EQUIP_FAX	Númerica		
TP_CATEGORIA_ESCOLA_PRIVADA	Catégorica	Baixo preenchimento - < 30% (13)	
IN_CONVENIADA_PP	Indicadora		
TP_CONVENIO_PODER_PUBLICO	Catégorica		

38 APÊNDICE B. PRÉ-PROCESSAMENTO DAS VARIÁVEIS DE ESCOLA DO CENSO ESCOLAR

IN_MANT_ESCOLA_PRIVADA_EMP	Indicadora	
IN_MANT_ESCOLA_PRIVADA_ONG	Indicadora	
IN_MANT_ESCOLA_PRIVADA_SIND	Indicadora	
IN_MANT_ESCOLA_PRIVADA_SIST_S	Indicadora	
IN_MANT_ESCOLA_PRIVADA_S_FINS	Indicadora	
CO_ESCOLA_SEDE_VINCULADA	Categórica	
CO_IES_OFERTANTE	Categórica	
TP_OCUPACAO_GALPAO	Categórica	
TP_INDIGENA_LINGUA	Categórica	
CO_LINGUA_INDIGENA	Categórica	
NU_ANO_CENSO	Numérica	Baixa variação - > 80% concentrado (122)
TP_SITUACAO_FUNCIONAMENTO	Categórica	
DT_ANO_LETIVO_INICIO	Data	
DT_ANO_LETIVO_TERMINO	Data	
CO_REGIAO	Categórica	
CO_UF	Categórica	
TP_DEPENDENCIA	Categórica	
TP_LOCALIZACAO	Categórica	
TP_REGULAMENTACAO	Categórica	
IN_LOCAL_FUNC_PREDIO_ESCOLAR	Indicadora	
TP_OCUPACAO_PREDIO_ESCOLAR	Categórica	
IN_LOCAL_FUNC_SALAS_EMPRESA	Indicadora	
IN_LOCAL_FUNC_SOCIOEDUCATIVO	Indicadora	
IN_LOCAL_FUNC_UNID_PRISIONAL	Indicadora	
IN_LOCAL_FUNC_PRISIONAL_SOCIO	Indicadora	
IN_LOCAL_FUNC_TEMPLO_IGREJA	Indicadora	
IN_LOCAL_FUNC_CASA_PROFESSOR	Indicadora	
IN_LOCAL_FUNC_GALPAO	Indicadora	
IN_LOCAL_FUNC_SALAS_OUTRA_ESC	Indicadora	
IN_LOCAL_FUNC_OUTROS	Indicadora	
IN_PREDIO_COMPARTILHADO	Indicadora	
IN_AGUA_FILTRADA	Indicadora	
IN_AGUA_REDE_PUBLICA	Indicadora	
IN_AGUA_POCO_ARTESIANO	Indicadora	
IN_AGUA_FONTE_RIO	Indicadora	
IN_AGUA_INEXISTENTE	Indicadora	
IN_ENERGIA_REDE_PUBLICA	Indicadora	
IN_ENERGIA_GERADOR	Indicadora	
IN_ENERGIA_OUTROS	Indicadora	
IN_ENERGIA_INEXISTENTE	Indicadora	
IN_ESGOTO_REDE_PUBLICA	Indicadora	
IN_ESGOTO_INEXISTENTE	Indicadora	
IN_LIXO_COLETA_PERIODICA	Indicadora	
IN_LIXO_QUEIMA	Indicadora	
IN_LIXO_JOGA_OUTRA_AREA	Indicadora	
IN_LIXO_RECICLA	Indicadora	
IN_LIXO_ENTERRA	Indicadora	
IN_LIXO_OUTROS	Indicadora	
IN_SALA_DIRETORIA	Indicadora	
IN_SALA_PROFESSOR	Indicadora	
IN_LABORATORIO_INFORMATICA	Indicadora	
IN_SALA_ATENDIMENTO_ESPECIAL	Indicadora	
IN_QUADRA_ESPORTES_COBERTA	Indicadora	
IN_QUADRA_ESPORTES	Indicadora	
IN_COZINHA	Indicadora	
IN_BIBLIOTECA	Indicadora	
IN_SALA_LEITURA	Indicadora	
IN_BIBLIOTECA_SALA_LEITURA	Indicadora	
IN_PARQUE_INFANTIL	Indicadora	
IN_BERCARIO	Indicadora	
IN_BANHEIRO_DENTRO_PREDIO	Indicadora	
IN_BANHEIRO_PNE	Indicadora	
IN_DEPENDENCIAS_PNE	Indicadora	
IN_SECRETARIA	Indicadora	
IN_BANHEIRO_CHUVEIRO	Indicadora	
IN_REFEITORIO	Indicadora	
IN_DESPENSA	Indicadora	
IN_ALMOXARIFADO	Indicadora	
IN_PATIO_COBERTO	Indicadora	
IN_PATIO_DESCOBERTO	Indicadora	

IN_AREA_VERDE	Indicadora	
IN_DEPENDENCIAS_OUTRAS	Indicadora	
IN_EQUIP_TV	Indicadora	
IN_EQUIP_DVD	Indicadora	
IN_EQUIP_COPIADORA	Indicadora	
IN_EQUIP_IMPRESSORA_MULT	Indicadora	
IN_EQUIP_SOM	Indicadora	
IN_EQUIP_MULTIMIDIA	Indicadora	
IN_EQUIP_FOTO	Indicadora	
IN_COMPUTADOR	Indicadora	
IN_INTERNET	Indicadora	
IN_BANDA_LARGA	Indicadora	
IN_ALIMENTACAO	Indicadora	
TP_AEE	Catagórica	
TP_ATIVIDADE_COMPLEMENTAR	Catagórica	
IN_FUNDAMENTAL_CICLOS	Indicadora	
TP_LOCALIZACAO_DIFERENCIADA	Catagórica	
IN_MATERIAL_ESP_QUILOMBOLA	Indicadora	
IN_MATERIAL_ESP_INDIGENA	Indicadora	
IN_MATERIAL_ESP_NAO_UTILIZA	Indicadora	
IN_EDUCACAO_INDIGENA	Indicadora	
IN_BRASIL_ALFABETIZADO	Indicadora	
IN_FINAL_SEMANA	Indicadora	
IN_FORMACAO_ALTERNANCIA	Indicadora	
IN_MEDIACAO_PRESENCIAL	Indicadora	
IN_MEDIACAO_SEMIPRESENCIAL	Indicadora	
IN_MEDIACAO_EAD	Indicadora	
IN_ESPECIAL_EXCLUSIVA	Indicadora	
IN_REGULAR	Indicadora	
IN_PROFISSIONALIZANTE	Indicadora	
IN_COMUM_CRECHE	Indicadora	
IN_COMUM_PRE	Indicadora	
IN_COMUM_FUND_AI	Indicadora	
IN_COMUM_FUND_AF	Indicadora	
IN_COMUM_MEDIO_MEDIO	Indicadora	
IN_COMUM_MEDIO_INTEGRADO	Indicadora	
IN_COMUM_MEDIO_NORMAL	Indicadora	
IN_ESP_EXCLUSIVA_CRECHE	Indicadora	
IN_ESP_EXCLUSIVA_PRE	Indicadora	
IN_ESP_EXCLUSIVA_FUND_AI	Indicadora	
IN_ESP_EXCLUSIVA_FUND_AF	Indicadora	
IN_ESP_EXCLUSIVA_MEDIO_MEDIO	Indicadora	
IN_ESP_EXCLUSIVA_MEDIO_INTEGR	Indicadora	
IN_ESP_EXCLUSIVA_MEDIO_NORMAL	Indicadora	
IN_COMUM_EJA_FUND	Indicadora	
IN_COMUM_EJA_PROF	Indicadora	
IN_ESP_EXCLUSIVA_EJA_FUND	Indicadora	
IN_ESP_EXCLUSIVA_EJA_MEDIO	Indicadora	
IN_ESP_EXCLUSIVA_EJA_PROF	Indicadora	
IN_COMUM_PROF	Indicadora	
IN_ESP_EXCLUSIVA_PROF	Indicadora	
INSE_NIVEL	Indicadora	
<hr/>		
IN_ESGOTO_FOSSA	Indicadora	Alta correlação - > 80% (10)
IN_BANHEIRO_EI	Indicadora	
IN_ALOJAM_PROFESSOR	Indicadora	
IN_LAVANDERIA	Indicadora	
NU_SALAS_EXISTENTES	Numérica	
NU_SALAS_UTILIZADAS	Numérica	
NU_EQUIP_IMPRESSORA_MULT	Numérica	
NU_COMPUTADOR	Numérica	
NU_COMP_ADMINISTRATIVO	Numérica	
IN_COMUM_EJA_MEDIO	Indicadora	

Referências Bibliográficas

- [1] Cidade Escola Aprendiz. *PROJETO ALUNO PRESENTE: Relatório 2013-2016*. 2017.
- [2] Fernando de Holanda Barbosa Filho and Samuel Pessôa. Retorno da educação no brasil. 2008.
- [3] RP Barros et al. Políticas públicas para a redução do abandono e da evasão escolar de jovens. *Página na internet, Fundação Brava, Insper, Instituto Unibanco e Instituto Ayrton Senna*, 2017.
- [4] Camila Bezerra, Ricardo Scholz, Paulo Adeodato, Tarcísio Lucas, and Itacira Ataíde. Evasão escolar: Aplicando mineração de dados para identificar variáveis relevantes. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 27, page 1096, 2016.
- [5] BRASIL. Lei nº 13.005, de 25 de jun. de 2014. Disponível em: http://www.planalto.gov.br/ccivil/_03/_ato2011-2014/2014/lei/l13005.htm. Acesso em: 26 jun. 2019.
- [6] Kennet Calixto, Caetano Segundo, and Renê Pereira de Gusmão. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1447, 2017.
- [7] Antonio Daniel Ricardo Engracia Caluz. *O papel das habilidades socioemocionais no fluxo escolar: uma análise do Ensino Médio brasileiro*. PhD thesis, Universidade de São Paulo, 2018.
- [8] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [9] Francisco HG Ferreira and Julie A Litchfield. Education or inflation?: the micro and macro-economics of the brazilian income distribution during 1981-1995. *Cuadernos de Economía*, 38(114):209–238, 2001.
- [10] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.
- [11] Harvey Goldstein, William Browne, and Jon Rasbash. Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences*, 1(4):223–231, 2002.
- [12] Ahlem Hajjem. Mixed effects trees and forests for clustered data. *Montreal: HEC Montreal, Department of Management Sciences*, 2010.
- [13] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

- [14] M Neri, E Gustafsson-Wright, G Seadlaceck, Daniela Costa, and Alexandre Pinto. The effects of idiosyncratic shocks to father's income on child labor, school drop-outs and repetition rates in brazil. *Anais do XXII Encontro Brasileiro de Econometria, SBE*, 2000.
- [15] Todos pela Educação. Quatro em cada 10 jovens de 19 anos ainda não concluíram o ensino médio. Disponível em: <https://www.todospelaeducacao.org.br/conteudo/quatro-em-cada-10-jovens-de-19-anos-ainda-nao-concluíram-o-ensino-medio>. Acesso em: 26 jun. 2019.
- [16] Maitê Rimekká Shirasu and Ronaldo De Albuquerque E Arraes. Determinantes Da Evasão E Repetência Escolar. Anais do XLIII Encontro Nacional de Economia 219, ANPEC - Associação Nacional dos Centros de Pós-Graduação em Economia, 2016.
- [17] UNICEF. *PANORAMA DA DISTORÇÃO IDADE-SÉRIE NO BRASIL*. 2018.