Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes

Camila Bezerra, Ricardo Scholz, Paulo Adeodato, Tarcísio Pontes e Itacira Silva

Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil

{cbs,reps,pjla,tdpl,ias3}@cin.ufpe.br

Abstract. *School dropout in Brazilian public education is a huge problem. One in four Brazilians leave school prematurely, before completing high school. This work deals with part of the problem, analyzing school dropout in the last year of Middle School on public schools of Pernambuco/Brazil, with data collected from the official National School Census, between 2011 and 2012. Decision Trees, Rules Induction and Logistic Regression were the Knowledge extraction techniques applied to identify the profile of a dropout student and estimate the propensity for that. Results show that age, classes' shift and geographic region strongly influence dropout.*

Resumo. *A evasão escolar na educação pública brasileira é um problema de grandes proporções. Um em cada quatro brasileiros deixa a escola prematuramente, antes de terminar o ensino médio. Este trabalho faz um recorte do problema, ao analisar a evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais e municipais do estado de Pernambuco, com base nos dados dos Censos Escolares 2011 e 2012. Árvore de Decisão, Indução de Regras e Regressão Logística foram as técnicas para extração de conhecimento aplicadas visando a identificar o perfil do aluno evasor e estimar a propensão à evasão. Os resultados mostraram que fatores como idade, turno das aulas e região geográfica das escolas influenciam fortemente a evasão.*

1. Introdução

Em 2012, o Brasil teve a terceira maior taxa de evasão escolar entre os cem países avaliados pelo Programa das Nações Unidas para o Desenvolvimento [UOL 2015]. Com uma taxa de evasão de 24,3%, o país destoa negativamente de vizinhos como Chile (2,6%) e Argentina (6,2%). O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) é o responsável por produzir e disseminar informações sobre a educação no Brasil em todos os níveis, desde o ensino infantil, até o superior [INEP 2015]. Segundo o Inep, 9,2% dos alunos do ensino médio abandonaram as escolas públicas brasileiras durante o ano letivo de 2013. No estado de Pernambuco, esse abandono foi de 5,2%, a menor taxa registrada no país naquele ano. Vale destacar que, embora o abandono tenha ligação com a evasão escolar, eles são conceitos diferentes. Pelos critérios do Inep, o abandono escolar se caracteriza quando o aluno deixa de frequentar as aulas e compromete o ano. Já a evasão ocorre quando o aluno abandona os

DOI: 10.5753/cbie.sbie.2016.1096

estudos e não retorna no ano seguinte [BRASIL 2015]. De acordo com a base de dados do Censo Escolar, no estado de Pernambuco, dos 118.755 alunos que concluíram o ensino fundamental em 2011, nas escolas públicas estaduais e municipais, 18.528 não se matricularam no ano seguinte, o que corresponde a uma evasão de aproximadamente 15,6%.

Os estudos sobre evasão escolar no Brasil costumam considerar três principais perspectivas do problema [Vieira 2007]: (1) o indivíduo, (2) a escola, e (3) o sistema de ensino. A primeira perspectiva diz respeito às condições de vida do aluno no seio familiar e, portanto, privado. Nesse âmbito, o Estado, em regra, só consegue agir de forma indireta como, por exemplo, mediante políticas públicas de distribuição de renda que exigem como contrapartida a frequência escolar dos menores de idade. Já os sujeitos das outras duas perspectivas, a escola e o sistema de ensino, são muito mais acessíveis à interferência direta do Estado, focada no problema específico da evasão escolar, de forma mensurável, objetiva e organizada. Ao entender o perfil da matrícula que ocasiona uma futura evasão, as políticas públicas podem focar nos alunos que estejam no grupo de risco desde o início do ano, de forma a reverter probabilidades.

Este trabalho faz um recorte do problema, investigando os fatores mais preponderantes na evasão de alunos que estão concluindo o ensino fundamental em escolas públicas estaduais ou municipais do estado de Pernambuco. O objetivo é trazer informações que contribuam para o desenvolvimento de políticas públicas visando a reduzir a evasão escolar nesse momento de transição da formação escolar.

Diante dessa perspectiva, o grão de decisão considerado foi a matrícula. Observe-se que um determinado aluno pode possuir mais de uma matrícula, no mesmo ano letivo. Entretanto, uma matrícula só é considerada como evasão se o aluno que a realizou não se matriculou no ano seguinte, em qualquer série, no estado de Pernambuco. A escolha do grão matrícula, em lugar do grão aluno, deu-se de forma a evitar a necessidade de agregação dos dados de múltiplas matriculas para determinados alunos.

Este estudo baseou-se em informações dos anos 2011 e 2012 do Censo Escolar [INEP 2015], um levantamento anual realizado pelo Inep, com centenas de variáveis descrevendo escolas, turmas, docentes e matrículas nos ensinos infantil, fundamental e médio, em todas as escolas do país. No registro de matrículas no Inep, cada aluno possui um identificador único que o acompanha ao longo dos anos. Dessa forma, é possível saber se um aluno que cursou o último ano do ensino fundamental, em 2011, por exemplo, realizou matrícula em alguma escola em 2012, e em que série.

O Censo Escolar de 2011 foi utilizado para construção dos modelos que buscaram generalizar o perfil dos alunos que não se matricularam no ano de 2012. O conhecimento foi extraído dos dados por meio de Árvore de Decisão, Indução de Regras e Regressão Logística, como técnicas de inteligência computacional cujo resultado foi testado em amostras de dados estatisticamente independentes das utilizadas na modelagem. Para a realização do projeto, seguiu-se a metodologia CRISP-DM, a mais usada em projetos de mineração de dados [Shearer 2000].

Na próxima seção, a base de dados e sua preparação para as etapas de extração de conhecimento são descritas em maiores detalhes. A terceira seção detalha cada uma das técnicas de inteligência computacional utilizadas. Na seção 4, realiza-se a análise

dos resultados encontrados. A quinta seção traz conclusões sobre a pesquisa, suas limitações e indica possíveis trabalhos futuros.

2. Descrição da Base de Dados e do Pré-Processamento

O Inep [INEP 2015] disponibiliza publicamente, desde 1995, em formato texto, os dados do Censo Escolar. Este estudo utilizou a base de dados de 2011 para extração de conhecimento e caracterizou a evasão com base nos dados do Censo de 2012.

O objetivo da fase de análise e preparação dos dados foi excluir os registros que não diziam respeito às perguntas de pesquisa, suprimir os campos irrelevantes ou redundantes e os registros inconsistentes, complementar campos incompletos (*missing values*) e apagar registros com valores muito dissonantes (*outliers*). Toda a etapa de preparação dos dados foi realizada em linguagem SQL, no SQLServer.

Os dados do Censo Escolar [INEP 2015] são divididos em quatro entidades: Matrícula (61 variáveis), Escola (123 variáveis), Turma (53 variáveis) e Docente (93 variáveis), totalizando 330 variáveis. A mineração dos dados ocorreu no grão "Matrícula". A base de dados original, do Censo Escolar 2011, continha informações de 2.699.350 matrículas das escolas públicas e privadas em Pernambuco, em todas as séries dos ensinos fundamental e médio.

A definição do escopo foi realizada pela filtragem para remoção dos registros que não estavam dentro do grupo de interesse. Nessa etapa, foram mantidas apenas as matrículas em escolas públicas estaduais e municipais do estado de Pernambuco, nas séries concluintes do ensino fundamental, excetuando-se a Educação de Jovens e Adultos e as modalidades de educação especial. Também foram excluídas matrículas de escolas que não haviam preenchido completamente os Censos Escolares dos anos 2011 ou 2012.

A caracterização da evasão (variável-alvo) foi definida pela ausência da matrícula de 2011 no ano de 2012, em qualquer escola do estado de Pernambuco, mesmo nas privadas, fora do escopo definido. Portanto, alunos que migraram e se matricularam em escolas de outros estados aparecerão erroneamente como evadidos mas esse contingente representa muito baixo percentual da população-alvo.

Posteriormente, as variáveis redundantes ou irrelevantes foram excluídas e analisou-se a base de dados em busca de registros com valores não preenchidos (*missing values*). Para as variáveis nessa situação, os registros foram preenchidos utilizando a moda e a média de cada atributo categórico e contínuo, respectivamente. Uma vez trabalhando no grão matrícula, foi necessário agregar os dados da tabela de docentes, de forma que a informação de vários docentes pudesse ser representada diretamente na tabela de matrículas.

Por fim, a base de dados resultante passou a apresentar 180 variáveis, sendo 142 delas categóricas e 54 numéricas. Restaram 118.755 matrículas, sendo 18.528 de alunos evasores (aproximadamente 15.6% do total), em 7.754 escolas, envolvendo 28.010 docentes e 3.673 turmas. Os dados foram particionados de forma estratificada pela variável-alvo em dois grupos: uma base de treinamento, com 79.209 registros (66,7% do total), e uma base de testes, com os 39.545 registros restantes (33,3% do total).

3. Mineração de Dados

3.1. Árvore de Decisão

A representação produzida pela Árvore de Decisão permite uma maior interpretabilidade das informações por seres humanos, particionando o espaço de entrada de maneira a maximizar o ganho de informação em relação à variável alvo, ponderando pela probabilidade de ocorrência de cada situação. A alta legibilidade da saída produzida pelas Árvores de Decisão facilita a validação por um especialista no domínio. Sendo assim, a aplicação dessa técnica é importante para o entendimento do fenômeno estudado, bem como a identificação de variáveis com maior poder de separação dos dados.

Independente de o método/algoritmo de geração de regras ser árvore de decisão ou indução, a qualidade das regras produzidas é avaliada por métricas clássicas. Neste trabalho estamos interessados nas regras de classificação; aquelas que têm a variável-alvo como consequente (aparece após o "então" da regra). Para regras de classificação, tipicamente, são medidos o suporte, a confiança e o *lift*. O suporte é a fração dos exemplos da amostra contidos no hipercubo definido pelo antecedente da regra (condições que aparecem entre o "se" e o "então" da regra). A confiança é a fração de exemplos da classe-alvo dentre os exemplos contidos no hipercubo da regra. A métrica de qualidade mais importante da regra é medida pelo *lift* que é a razão entre a confiança da regra e a confiança da amostra de dados (constante, de 0,156, neste artigo).

Neste trabalho, utilizou-se o método CHi-squared Automatic Interaction Detection (CHAID) Exaustivo, disponível no software SPSS [IBM 2015]. Para as variáveis categóricas com mais de duas categorias, o método examina todas as possíveis partições dos dados, buscando a partição ótima, e eventualmente agrupando partições, de acordo com a significância [Kass 1980][Biggs et al. 1991].

A parametrização foi realizada buscando balancear a relação custo-benefício entre a profundidade da árvore e a relevância das subdivisões, em contraponto à possibilidade de encontrar "pepitas de conhecimento", que ocorrem apenas em situações menos frequentes. A poda da árvore foi realizada por duas estratégias: limitando-se a profundidade máxima a cinco níveis, e limitando-se a representatividade dos nós pais e filhos, respectivamente, a 6.000 exemplos (aproximadamente 5% da base) e a 1.200 exemplos (aproximadamente 1% da base). Embora essa abordagem possa ter implicado em perda de acurácia do modelo gerado, evitou-se a exibição de nós com representatividade muito baixa nas folhas da árvore, ou geração de uma árvore muito profunda, o que dificultaria a compreensão das regras por especialistas.

Quanto aos parâmetros que influenciam no fatiamento de valores das variáveis categóricas, utilizou-se o nível de significância de 5% para permitir a separação. A respeito dos critérios de parada, ajustou-se o número máximo de iterações sem que houvesse modificação do modelo para 1.200 (aproximadamente 1% dos dados). Além disso, a alteração mínima na frequência de células esperadas foi configurada para 0,1%. Os valores de significância foram ajustados pelo método de Bonferroni, e permitiram-se novas divisões de categorias previamente agrupadas dentro de um nó. A Figura 1 mostra os principais ramos da árvore.

No primeiro nível, a variável de separação dos dados foi a idade dos alunos (nó 0). Percebe-se que, entre os alunos mais jovens, até cerca de 16 anos, a evasão não

supera 7,8% (nós 1, 2 e 3), muito abaixo da taxa média de 15,6%. Já entre os alunos mais velhos, a taxa extrapola para 46,8% dos 19 anos em diante (nó 9). Claramente, os alunos mais velhos precisam de um acompanhamento maior, para evitar a evasão.

No segundo nível da árvore, a influência do tempo de permanência diária dos alunos na escola se destaca. Para a faixa-etária de 16 a 17 anos (nós 5, 6 e 7 da árvore), escolas em que a duração das aulas é de até quatro horas têm evasão muito superior. Ou seja, escolas em tempo integral têm baixo nível de evasão, para essa faixa-etária.

A árvore deixa claro que o grande problema de evasão escolar ocorre com os alunos de maior faixa-etária (nó 9). Isso deixa evidente que outras demandas, incluindo a pressão para trabalhar visando à sua sustentação interferem muito na continuidade dos estudos para os alunos mais velhos.

A área geográfica em que se encontra a escola também aparece como fator relevante, embora com menores ganhos de informação. Observa-se, tanto no segundo quanto no terceiro nível, que para grupos específicos, o Órgão Regional Inep é um fator

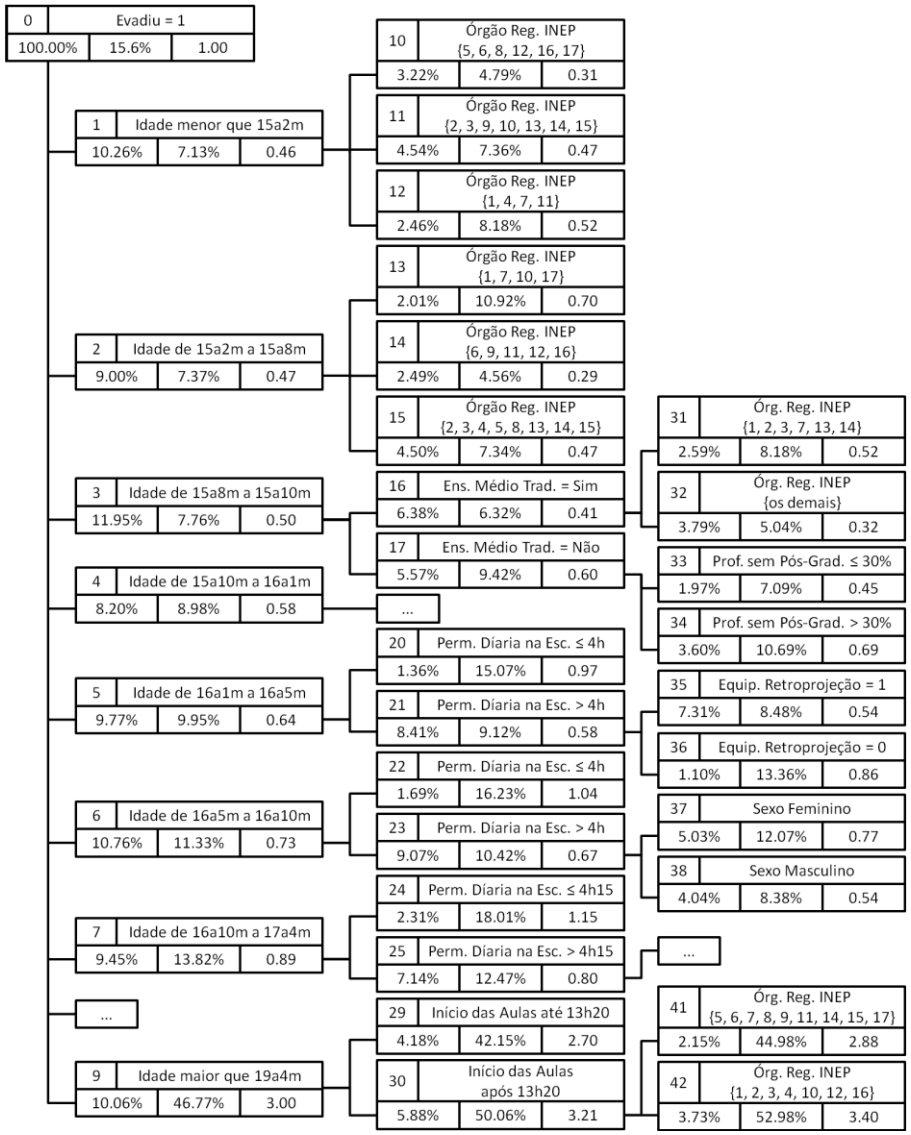

Figura 1. Árvore de Decisão: na linha inferior de cada célula, da esquerda para a direita, suporte, confiança e *lift*.

discriminante (nós 10 a 15, e nós 31 e 32). Embora não tenha sido possível encontrar a listagem dos códigos de Órgãos Regionais do Inep, alguns códigos, como o 1 e o 7, tendem a aparecer sempre nos ramos que possuem maior evasão, enquanto outros, como o 6, o 12 e o 16, aparecem repetidamente entre os que geram ramos com menor evasão. É importante associar esses Órgãos Regionais às condições de infraestrutura e aos dados demográficos das suas regiões no estado de Pernambuco.

Por fim, no terceiro nível da árvore, para os alunos na faixa-etária dos 15 anos que cursam o ensino médio não tradicional (técnico, magistério, etc), a quantidade de professores que não possuem pós-graduação aparece como fator relevante no percentual de evasão. Assim, a formação dos docentes acaba por influenciar a evasão, ainda que menos que outras variáveis. Supõe-se que professores melhor formados proporcionem aulas de melhor qualidade e isso estimule os alunos a permanecerem em sala de aula.

3.2. Indução de Regras

A indução de regras é uma das técnicas mais importantes na mineração de dados [Maimon and Rokach 2005]. Os algoritmos dessa família geram regras de associação do tipo "se-então", facilmente entendidas por um humano. Neste trabalho, dentre as regras geradas selecionamos apenas as regras de classificação (as que têm a variável-alvo como consequente). Essas regras são redundantes e não cobrem todo o espaço de entrada [da Silva and Adeodato 2012]. Este estudo utilizou o algoritmo de indução de regras *a priori* [Borgelt and Kruse 2002], no software estatístico R [R 2015]. Um dos principais parâmetros do algoritmo *a priori* é o suporte mínimo para a regra não ser descartada que neste trabalho foi de 0,01. Outro parâmetro importante é o número máximo de cláusulas da regra que foi limitado a três atributos, uma vez que o seu aumento dificulta a interpretabilidade das regras por humanos e aumenta exponencialmente o custo computacional para execução do algoritmo.

Considerando as regras com o suporte mínimo de 1%, a qualidade das regras de classificação deve ser medida em relação ao desvio da média de 15,6% de confiança, representado pelo *lift* em relação ao valor 1. O interesse é identificar os nichos que tenham ou alta taxa de evasão para combater essas condições ou baixa taxa de evasão para poder replicar tais condições. Neste trabalho, dado que a taxa de evasores é de 15,6%, o *lift* máximo para o rótulo evasor é 6,41 (=1/0,156). A Tabela 1 mostra as cinco regras de maior e menor *lift*.

Segundo a indução de regras, o turno e a idade dos alunos são os atributos mais determinantes para a evasão escolar, uma vez que aparecem em todas as regras de maior e menor *lift*. Alunos que estudam à tarde ou à noite e possuem mais de 16 anos e meio de idade chegam a apresentar um *lift* de 2,18, corroborando o conhecimento explicitado pela Árvore de Decisão. Já a presença de atributos como o número de salas utilizadas ou o número de salas existentes na escola, com *lifts* próximos a 2,00, indicam que escolas de maior porte tendem a apresentar maior evasão. Isso é corroborado pelas regras de baixo *lift*, em que escolas com menor número de funcionários por aluno matriculado obtiveram o menor *lift* (0,47). Outro atributo relevante foi a existência de ensino religioso: quando a disciplina não é oferecida ou não há professor para ela, os valores de *lift* chegam muito próximos de 2,00, indicando que o ensino religioso tem relação com a manutenção dos alunos na escola.

Entre as regras com menor *lift*, o número de alunos por sala é um indicativo de que turmas com menos alunos tendem a apresentar taxas de evasão mais baixas. Além disso, o ensino médio tradicional (*lift* de 0,43), a presença de quadra na escola (*lift* de 0,46) e a administração da escola pelo estado (*lift* de 0,45) também são itens que influenciam positivamente, evitando a evasão dos alunos.

Tabela 1. Regras de maior e menor *lift*, no universo com 15,6% de evasão.

Condição 1	Condição 2	Sup.	Conf.	Lift
Turno da tarde ou turno da noite	Aluno com mais de 16 anos e meio	5,0%	34,0%	2,18
Turno da tarde ou turno da noite	A partir de 13 salas de aula existentes na escola	2,9%	31,0%	1,98
Turno da tarde ou turno da noite	Turma não possui ensino religioso	3,4%	31,0%	1,98
Aluno com menos de 16 anos e meio	Número de funcionários é reduzido (um para cada 15,5 ou mais alunos)	1,8%	7,4%	0,47
Aluno com menos de 16 anos e meio	Escola possui quadra descoberta	1,1%	7,2%	0,46
Aluno com menos de 16 anos e meio	Escola apresenta média de menos de 70 alunos por sala utilizada (somando-se os turnos)	1,7%	7,2%	0,46
Aluno com menos de 16 anos e meio	Escola é administrada pelo estado	2,1%	7,1%	0,45
Aluno com menos de 16 anos e meio	Regimento de ensino médio padrão (em três anos, não seriados e não profissionalizante)	1,8%	6,8%	0,43

3.3. Regressão Logística

Complementando os resultados obtidos com as demais técnicas, a Regressão Logística é um modelo linear generalizado que relaciona as variáveis de entrada à variável alvo. A saída desse modelo permite quantificar quais variáveis explicativas são consideradas mais relevantes para a definição da variável alvo [Hosmer and Lemeshow 2000].

A variável dependente binária é definida da seguinte forma: $y_i = 1$, se o aluno evadiu, e $y_i = 0$, se o aluno não evadiu. Seja $i = P\ (y_i = 1)$ a probabilidade de o *i*-ésimo aluno a evadir, o modelo de Regressão Logística é representado pela expressão a seguir:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Onde β_n são parâmetros desconhecidos a serem estimados. Nesse caso, a probabilidade de o *i*-ésimo aluno evadir é dada por:

$$P\left(Evadir_i \middle| x_{i1}, x_{i2}, \ldots, x_{ip}\right) = \pi_i = \frac{1}{1 + exp\left\{-\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)\right\}}$$

A regressão logística foi executada no software R [R 2015], com o modelo "*Backward Stepwise*", o qual incorpora inicialmente todas as variáveis e depois, por etapas, retira ou não cada variável do modelo.

Das 180 variáveis explicativas, 125 foram preservadas no modelo ao nível de significância $p < 0,01$. A Tabela 2 ilustra aquelas de maior influência, seja positiva ou negativamente, exibindo os seus coeficientes β_n.

Com relação aos docentes, os fatores mais fortes são a dedicação exclusiva (média de escolas em que o docente trabalha) e o número de docentes pós-graduados.

Itens estruturais da escola mostraram-se importantes nos níveis de evasão: presença de quadras esportivas, laboratórios de ciências e tamanho da escola (pela quantidade de salas de aula disponíveis). Por fim, assim como na árvore de decisão e na indução de regras, a idade do aluno e o turno de estudo foram considerados variáveis com forte influência na evasão dos alunos.

Tabela 2. Regressão Logística

Variável	β
Escola possui educação infantil especial	0,49
Percentual dos professores da turma sem pós-graduação	0,25
Escola possui atendimento educacional especializado	0,15
Regimento de ensino: Médio (outros tipos: Integrado, Normal/Magistério, Profissional)	0,12
Média do número de escolas em que os docentes da turma trabalham	0,10
Laboratório de ciências	0,10
Local de funcionamento da escola - Salas em outra escola	-0,10
Abastecimento de água – Poço artesiano	-0,17
Local de funcionamento da escola – Unidade de Internação/Prisional	-0,33
Percentual de professores da turma que moram em PE	-0,84

4. Análise dos Resultados

Neste trabalho, consideramos dois métodos para teste de desempenho e mensuração da qualidade dos modelos de classificação: máxima distância de *Kolmogorov-Smirnov* (KS, Máx_KS) e área sob a curva ROC (*Receiver Operating Characteristic*, AUC_ROC). Essas medidas foram computadas sobre a amostra de teste, contendo 33,3% da massa de dados total, extraída por amostragem aleatória estratificada pela classe-alvo.

O teste de KS consiste em obter a máxima diferença entre distribuições acumuladas das amostras da classe alvo e da complementar [GREBIN 2012]. Quanto mais separadas estiverem as distribuições, melhor é o modelo de acordo com essa métrica que é restrita a apenas um ponto de potencial operação que, em geral não corresponde aos interesses do domínio. A Figura 2 mostra o KS máximo de 0,307 obtido pela regressão. Apesar da equivalência entre as curvas ROC e KS [Adeodato e Melo 2016], a métrica de área sob a curva ainda é mais difundida.

A curva ROC é um gráfico que mostra a taxa de verdadeiros positivos *versus* a taxa de falsos positivos. É uma representação gráfica da sensibilidade, em relação ao complemento da especificidade [Fawcett 2006]. Neste trabalho, sensibilidade corresponde à probabilidade de identificar os alunos propensos a evadir, e especificidade, à probabilidade de identificar os alunos não propensos a evadir.

A curva ROC ideal consiste em duas retas no eixo vertical, uma que vai do ponto (0,0) ao ponto (0,1), e outra que segue até o ponto (1,1). Isso indica que o classificador permite classificar toda a amostra corretamente. Assim, a taxa de verdadeiros positivos é igual a 1, e a taxa de falsos positivos é igual a 0 [Fawcett 2006]. A área sob a cur-

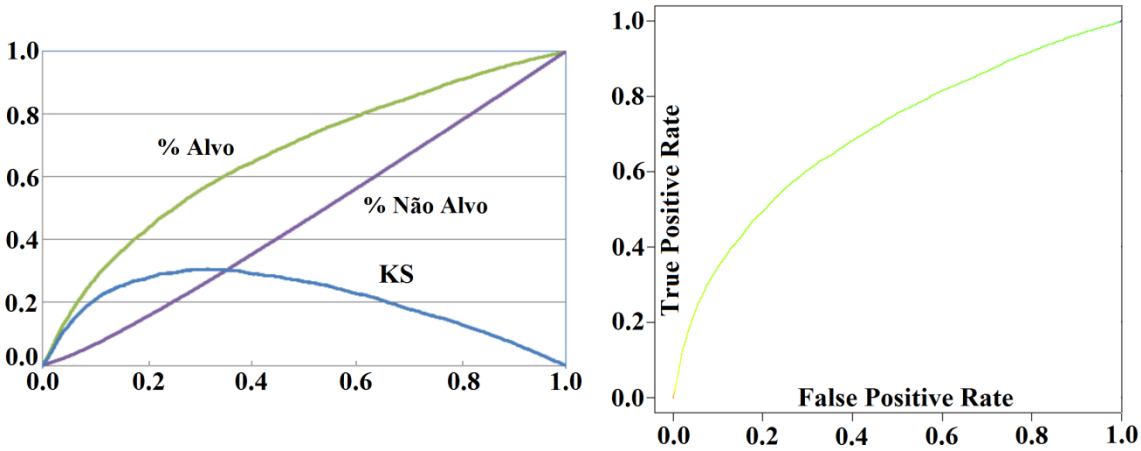

Figura 2. Da esquerda para a direita: curva KS e curva ROC.

va ROC é uma métrica de desempenho que integra toda a faixa de escore do classificador e quanto mais a curva se afasta da diagonal principal, melhor é o classificador. A Figura 2 mostra a curva ROC obtida pelo modelo de regressão logística, com área 0,69.

5. Conclusões

Este artigo abordou o problema da evasão das escolas municipais e estaduais de Pernambuco no último ano do ensino fundamental com base nas informações do Censo Escolar de 2011, considerando evasão a não-matrícula do aluno em escola do estado no ano seguinte, com o objetivo de descobrir fatores que levam estudantes a evadir. A partir do conhecimento desses fatores, as escolas podem criar mecanismos para diminuir a evasão.

A Árvore de Decisão apresentou regras em formato de decisão sequencial. A Indução de Regras explicitou o conhecimento em forma de regras "*se-então*", Ambas tiveram sua qualidade avaliada por meio das métricas Cobertura, Confiança e *Lift*. A Regressão Logística estimou a chance de evasão de cada aluno em uma amostra estatisticamente independente cuja qualidade foi medida pelo Máx_KS e AUC_ROC. As três técnicas produziram resultados consistentes entre si.

Verificou-se forte relação entre a evasão escolar e a idade dos alunos, o turno em que estudam, a estrutura e o tamanho da escola, o tempo de permanência em sala de aula, a existência de aulas de religião ou educação física ou a dedicação dos professores. O conhecimento extraído está pronto para ser embarcado em sistemas de suporte à decisão que utilizem a metodologia CRISP-DM [Shearer 2000], de domínio público.

A metodologia apresentada pode ser replicada a cada ano e incorporar dados de vários anos de histórico dos alunos, trazendo aspectos evolutivos ainda não considerados neste trabalho.

O presente trabalho ainda possui limitações. Uma delas é a granularidade de algumas variáveis: por exemplo, a idade deverá ser descrita em anos completos e o início das aulas deverá ser trocado pelo turno das aulas, em futuras simulações. Além disso, algumas simplificações podem ter afetado os resultados, como a decisão de considerar evasores os alunos que não se matricularam no estado de Pernambuco no ano seguinte. Ainda como trabalho futuro, seria interessante realizar um estudo comparativo entre os estados da federação com níveis de evasão similar, para identificar, por

exemplo, se os motivos são parecidos. Por fim, simulações com bases de dados mais recentes, ou incluindo um histórico mais longo dos alunos (o que possibilitaria acrescentar variáveis como quantidade de repetências em cada série, por exemplo) certamente aumentarão a confiabilidade dos resultados e a possibilidade de encontrar "pepitas de conhecimento".

7. Referências

Adeodato, P. J. L. and Melo, S.B. (2016). On the equivalence between Kolmogorov-Smirnov and ROC curve metrics for binary classification. Cornell University Library ARXIV, 2016arXiv160600496A, https://arxiv.org/abs/1606.00496.

Biggs, D., Ville, B. D., and Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18:49–62.

Borgelt, C. and Kruse, R. (2002). Induction of association rules: Apriori implementation. In *Proceedings of the 15th Conference on Computational Statistics*.

BRASIL (2015). Disponível em http://www.brasil.gov.br/educacao/2012/05/indice-de-abandono-escolar-e-tres-vezes-maior-no-6o-ano-do-ensino-fundamental. Acessado em 14/08/2016.

da Silva, H. R. B. and Adeodato, P. J. L. (2012). A data mining approach for preventing undergraduate students retention. In *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*, pages 1–8. IEEE.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley series in probability and statistics. John Wiley & Sons, Inc. A Wiley-Interscience Publication, New York, Chichester, Weinheim.

IBM (2015). IBM SPSS Software Trial Version. Disponível em http://www.ibm.com/software/ analytics/ spss/. Acessado em 14/08/2016.

INEP (2015). Disponível em http://portal.inep.gov.br/ . Acessado em 14/08/2016.

GREBIN. S. Z. (2012). "Combinação em série e em paralelo de modelos de redes neurais e regressão logística - um estudo de caso em crosselling," Trabalho de conclusão do curso de Bacharelado em Estatística, Universidade Federal do Rio Grande do Sul.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2):119–127.

Maimon, O. and Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4).

UOL (2015). http://educacao.uol.com.br/noticias/2013/03/14/brasil-tem-3-maior-taxa-de-evasao-escolar-entre-100-paises-diz-pnud.htm. Acessado em 14/08/2016.

R (2015). Disponível em http://www.r-project.org/. Acessado em 14/08/2016.

VIEIRA, S. L.; FARIAS, I. M. S. (2007) Política Educacional no Brasil: introdução histórica. Brasília: Liber Livro.