Análise de Classificadores para Predição de Evasão de Campi de uma Instituição de Ensino Federal

Fernando Wagner B. H. Filho[1], Tiago S. Vinuto[2], Bruno C. Leal[3]

[1]Instituto Federal do Ceará (IFCE)
Paracuru – CE – Brasil

[2]Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brasil

[3]Instituto Federal do Piauí (IFPI)
Floriano – PI – Brasil

fernando.wagner@ifce.edu.br, tiagosv@lia.ufc.br, brunoleal@ifpi.edu.br

Abstract. *Student dropout is a serious problem present in educational institutions around the world. There are a variety of factors that can trigger student dropouts and several actions have been taken to mitigate the dropout rates. However, the vast majority of these actions are reactive. This work presents an analysis of the student dropout problem in the most affected campuses of Federal Institute of Ceará. In this analysis, student dropout is defined as a classification problem, which aims to generate models for predicting student dropout using the leading machine learning classifiers. This approach allows the pedagogical team, teachers, and other institution sectors to make preventive decisions against student dropouts.*

Resumo. *A evasão escolar é um grave problema em instituições educacionais de todo o mundo. Inúmeros fatores podem desencadear evasões discentes e várias ações vêm sendo tomadas para de mitigar as taxas de evasão. Porém, a maioria dessas ações são reativas. Este trabalho apresenta uma análise da problemática da evasão nos campi mais afetados do Instituto Federal do Ceará. Nesta análise, a evasão discente é definida como um problema de classificação, cujo objetivo é gerar modelos para predizer a evasão discente utilizando classificadores de aprendizagem de máquina. Essa abordagem possibilita que a equipe pedagógica, professores e demais setores da instituição possam tomar decisões preventivas contra a evasão estudantil.*

1. Introdução

Nas últimas décadas a informática vem se destacando por prover ferramentas computacionais voltadas para a facilitação da resolução de problemas das diversas áreas de conhecimento. No que concerne à educação, soluções vêm sendo criadas para atuar em questões de ensino e aprendizagem, como a criação de objetos de aprendizagem e ambientes virtuais de aprendizagem (AVAs), viabilizando inclusive a prática da chamado ensino a distância (EaD) e o ensino remoto.

A tecnologia da informação tem aumentado sua atuação na área de gestão educacional com automatização de funções administrativas e, até mesmo pedagógicas, por

DOI: 10.5753/cbie.sbie.2020.1132

parte das instituições e/ou seus pares. Isso se dá através de sistemas de informações, além de outras ferramentas de armazenamento e gerenciamento de dados. A grande maioria dos sistemas automatiza e torna eficiente tarefas preexistentes. Porém, a análise dos dados, geralmente, ainda fica a cargo de indivíduos. E, devido ao grande volume de dados, torna-se quase impossível para uma pessoa inferir previamente possíveis problemas, e.g., a tendência de abandono de cada um dos estudantes de uma instituição de ensino (IE).

Nos últimos anos, a aprendizagem de máquina (do inglês, *Machine Learning*) vem ganhando destaque auxiliando em problemas relacionados à tomadas de decisão, fornecendo ferramentas e métodos que computam informações pirobalísticas para os mais diversos cenários que possam vir a acontecer com base em um conjunto de dados prévios. O presente trabalho aplica os principais modelos de aprendizagem de máquina para a predição de possíveis evasões discentes considerando dados dos campi com maior evasão do Instituto Federal do Ceará (IFCE). Os resultados de uma boa classificação podem auxiliar a gestão acadêmica de uma instituição no tratamento do problema da evasão, pois permite agir por meio de ações preventivas e tomadas de decisão a nível institucional.

O restante deste trabalho está organizado como a seguir. A Seção 2 apresenta os principais conceitos e definições acerca do problema da evasão discente. Logo a seguir, a Seção 3 aborda a área de aprendizagem de máquina com foco no problema de classificação de dados. Em seguida, a Seção 4 apresenta trabalhos relacionados na área. A Seção 5 detalha a metodologia aplicada no desenvolvimento do presente trabalho. Já na Seção 6 são expostos os resultados, além de discussões e novas questões instigadas a partir dos mesmos. Por fim, a Seção 7 apresenta as conclusões e trabalhos futuros.

2. O Problema da Evasão Discente

A evasão discente vem se destacando como um grave problema para instituições de ensino ao redor do mundo. Em [Silva Filho and Araújo 2017] define-se o problema da evasão como sendo o desligamento do aluno da instituição de ensino no qual o mesmo não retorna mais ao sistema escolar. Os mesmos autores ainda apontam que a evasão pode estar ligada a vários fatores de natureza pessoal ou institucional. Segundo [IFCE 2017], são exemplos de fatores institucionais a falta de infraestrutura como defasagens no acervo bibliográfico, ausência de locais adequados para estudo e convivência e ausência ou mal-estado de conservação dos equipamentos para práticas (como máquinas em laboratórios de informática). O mesmo documento destaca como exemplo de fatores de natureza pessoal a conjuntura sócio-econômica do país no momento, desemprego/busca por emprego, o baixo rendimento escolar e a não identificação com o curso escolhido.

Como exemplo de ações de combate à evasão, [BRASIL 2014] propõe algumas ações que visam diminuir os índices e consequências causados por este fenômeno. Dentre algumas das ações citadas, pode-se destacar a adequação dos horários de aula em consonância com o sistema de transporte público, acompanhamento de estudantes que apresentam recorrente falta de pontualidade e/ou assiduidade, e a frequente revisão de matrizes curriculares e duração dos cursos, além da eventual readequação de infra-estruturas porventura ausentes ou defasadas.

Ainda analisando o documento de [BRASIL 2014], percebe-se que muitas ações poderiam ser otimizadas caso houvesse uma identificação prévia daqueles alunos que possuem uma maior predisposição de evasão da instituição, como o acompanhamento mais

próximo por parte do docente, da equipe técnico-pedagógica da instituição, e o incentivo financeiro através da oferta de bolsas ou auxílios estudantis. Neste contexto, ferramentas computacionais da área de aprendizagem de máquina podem ser um mecanismo de identificação precoce destes casos e, portanto, um valioso instrumento de auxílio do setor pedagógico e outros setores de apoio ao estudante.

3. Classificação em Aprendizagem de Máquina

O aprendizado de máquina é um sub-campo da inteligência artificial que aborda a questão sobre como tornar as máquinas aptas a aprender [Rätsch 2004]. Ainda de acordo com [Géron 2019], a aprendizagem de máquina também pode ser vista como a "arte de programar um computador para aprender com os dados", que visa o desenvolvimento de técnicas e métodos para aperfeiçoamento da realização de uma tarefa por meio da interpretação dos dados fornecidos ao longo do tempo.

Em [Silveira and Bullock 2017] são mencionados exemplos de problemas que podem ser atacados com aprendizagem de máquina. Os autores citam o governo dos Estados Unidos (EUA) que utiliza algoritmos de classificação na tentativa de identificar padrões de transferência de fundos internacionais de lavagem de dinheiro do narcotráfico e prevenção de atentados. No varejo comercial, os autores apontam empresas, como a *Walmart*, que vêm construindo soluções baseadas em aprendizagem de máquina, a partir de dados de compras e preferências de clientes, para definir, por exemplo, a melhor organização de prateleiras e a previsão de necessidades de produtos e serviços por determinados clientes. Já na área da saúde, a investigação é realizada na tentativa de encontrar padrões entre doenças e perfis profissionais, socioculturais, hábitos pessoais e locais de moradia, com intuito de propor um melhor entendimento dessas doenças e seus tratamentos.

A aprendizagem de máquina, segundo [Bruce and Bruce 2019], pode ser dividida em aprendizado supervisionado e não supervisionado. No aprendizado supervisionado o objetivo é compreender a relação entre as entradas e saídas fornecidas, com intuito de classificar, ou rotular, uma determinada instância (imagem, produtos, documento, etc.). Para isso se utiliza de um conjunto de categorias pré-definidas com auxílio de um "supervisor" que sabe previamente qual é a resposta esperada dado uma entrada. Este tipo de aprendizagem é divido em: *(a)* regressão, cujo propósito é encontrar como uma variável evolui em relação a outras; e *(b)* classificação, que consiste em atribuir um rótulo para a saída a partir de determinada entrada. Já no aprendizado não supervisionado, o objetivo é agrupar elementos com características similares formando grupos (*clusters*). Neste tipo de aprendizado o algoritmo atua sem um supervisor e, além disso, não existe influência humana direta, assim o conhecimento se dá a partir do agrupamento dos dados baseados em suas similaridades.

Este trabalho utiliza a abordagem supervisionada do tipo classificação, que tem como objetivo determinar se uma instância pertence a uma ou mais classes [Sebastiani 2002]. Utilizamos e comparamos alguns dos principais classificadores existentes na literatura para definição de modelos e aplicação na predição de evasão discente.

4. Trabalhos Relacionados

Dado o aumento na utilização de tecnologias de aprendizagem de máquina nos mais diversos nichos, conforme apresentado na Seção 3, e o nível de relevância do problema da

evasão discente na educação, vários trabalhos abordam estes pontos considerando diversas situações e cenários. A Seção 4.1 aborda alguns destes trabalhos em âmbito nacional, enquanto que a Seção 4.2 explora trabalhos desenvolvidos internacionalmente.

4.1. Trabalhos Nacionais

No trabalho desenvolvido em [Rigo et al. 2014], os autores defendem a utilização de mecanismos de aprendizado de máquina para o diagnóstico e combate ao problema da evasão discente, destacando que estes devem ser acompanhados de ações pedagógicas para que haja um real aproveitamento de seus resultados.

O trabalho de [Amorim et al. 2008] utilizou classificadores baseados em árvore de decisão, máquina de vetores de suporte (SVM) e redes Bayesianas, por meio da ferramenta *Weka*[1], para predizer a evasão discentes em cinco cursos de graduação plena de uma Instituição de Ensino Superior (IES) particular no município de Campos dos Goytacazes-RJ. Como resultado, o estudo acabou obtendo uma acurácia máxima da ordem de 91%. No entanto, por não ser recente, o trabalho não considerou outras métricas de avaliação que hoje são frequentemente utilizadas, como precisão (Eq. 2) e revocação (Eq. 3), comprometendo a interpretação do resultado como um todo. Além disso, os classificadores utilizados foram restritos quando comparados com as opções existentes atualmente.

Em [Queiroga et al. 2017] a ferramenta *Weka* também foi utilizada para fornecer classificadores com o intuito de predizer alunos de cursos técnicos na modalidade a distância com alto risco de evasão. Para isso foram utilizados dados oriundos do quantitativo de interações no AVA adotado pela instituição do estudo de caso, bem como outras informações derivadas desse quantitativo. Tal estudo foi motivado pelos altos índices de evasão no nível e modalidade selecionados, e os resultados foram considerados relevantes em relação a outras abordagens existentes na literatura. A distinção deste trabalho para a contribuição aqui apresentada, se dá no tipo de dados utilizados e, apesar de bastante relevante, não se traça um paralelo direto.

No que concerne aos institutos federais, o trabalho descrito em [Saraiva et al. 2019] observou o considerável nível de evasão dos cursos técnicos de informática da instituição abordada no estudo de caso. O trabalho focou na construção de modelos para predição de alunos com tendência de evasão nestes cursos. Para isso, obtiveram e prepararam uma base com dados de alunos referentes ao período de 2008 a 2018, contendo mais de 700 registros. Como resultados, obtiveram níveis de acurácia na ordem de 97%. Acreditamos que o altíssimo nível de acurácia deste trabalho se deve à utilização de uma base de dados bem específica englobando alunos de um mesmo curso de um mesmo campus. O autor menciona este fato e destaca como trabalho futuro a ampliação de aplicação do trabalho. Por ser um trabalho recente, é sentida a ausência de outras métricas de análise, como precisão, revocação, especificidade ou medida-F, para uma análise mais detalhada do resultado.

4.2. Trabalhos Internacionais

Em âmbito internacional, o trabalho descrito em [Tan and Shao 2015] teve por objetivo definir modelos de predição de potenciais alunos em condição de evasão, considerando dados discentes da Universidade Aberta da China (*Open University of China*) – filial

[1]https://www.cs.waikato.ac.nz/ml/weka/

(campus) *Sichuan*. Em seus experimentos os autores utilizaram três classificadores baseados em árvores de decisão, redes neurais e redes Bayesianas, apresentando níveis razoáveis de medida-F (Eq. 5). O artigo conseguiu apresentar boas taxas de acerto de alunos não evadidos, mas para evadidos foi um pouco baixa, que foi justamente o que impactou na medida-F (máximo de 0.71).

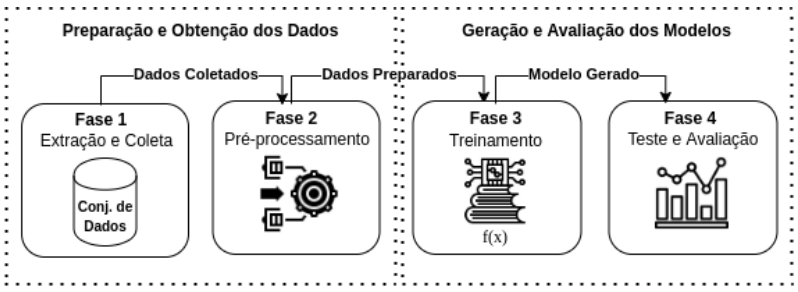
O trabalho de [Sansone 2019] combina teorias econométricas com algoritmos de aprendizagem de máquina para construção de modelos de predição para evasão do nono ano (*9th grade*) de uma escola de ensino médio norte-americanas (*high school*), mostrando que ambas as áreas podem cooperar entre si, obtendo resultados relevantes, em especial, quando da utilização de técnicas de SVM (*Support Vector Machine*). Apesar de obter métrica de revocação baixa (menor que 0.30), obteve bons níveis de acurácia (Eq. 1) e AUC (Seção 5.2).

Em [Chung and Lee 2019] foram utilizados dados do *National Education Information System* da Coreia do Sul, que integra dados de mais de 12.000 escolas espalhadas por todo o país. Deste contingente, extraiu-se uma amostra de aproximadamente 165.000 estudantes do ensino médio oriundos de 4 cidades, incluindo a capital Seoul. Através dos experimentos, algoritmos baseados em floresta randômica (*Random Forest*) tiveram um resultado considerado interessante aos propósitos do estudo. Obteve valores muito bons de curva ROC (Seção 5.2), bons valores de revocação (Eq. 3), especificidade (Eq. 4), porém não apresenta a precisão, que computada pela matriz de confusão (Seção 5.2) é muito baixa (0.11).

5. Metodologia Aplicada

Esta seção apresenta a metodologia aplicada ao longo do desenvolvimento deste trabalho que é resumida e ilustrada na Figura 1. Dentre os principais tópicos, estão a forma como os dados foram obtidos e preparados, a escolha e definição dos classificadores utilizados, e como os resultados foram gerados e avaliados.

Figura 1. Visão geral da metodologia utilizada



5.1. Obtenção e Preparação dos Dados

O IFCE, utilizado como estudo de caso neste trabalho, assim como várias instituições, sofre com problema de alta taxa evasão discente, problema que se agravou, ainda mais, ao longo da última década. Como forma de combate ao problema, uma importante iniciativa foi criada: a plataforma IFCE em Números [Pró-reitoria do IFCE 2017]. O intuito principal da plataforma é a publicização dos dados relacionados ao ensino dos campi de forma centralizada e fácil, não somente para a gestão da instituição, mas para com os demais membros de sua comunidade acadêmica. A plataforma também possui o importante

papel de fornecer uma visão sistêmica da instituição e seu contexto, de modo a auxiliar na elaboração e planejamento de ações de combate a evasão e retenção discente.

Fase 1: Extração e Coleta dos Dados. Os dados foram obtidos por meio da plataforma IFCE em Números, onde selecionou-se como estudo de caso os dados dos 5 campi com maior taxa de evasão discente. Foram, então, obtidos os dados de alunos agregados por campus, por curso e por período em formato de planilhas eletrônicas. A partir dessas planilhas, se fez possível a utilização de bibliotecas para extração, manipulação, processamento e posterior análise desses dados. Verificando os dados disponíveis, as seguintes variáveis foram selecionadas: *i)* é ingressante no período ou não; *ii)* encontra-se retido no curso ou não; *iii)* é cotista ou não; *iv)* a etnia do aluno; *v)* sexo; *vi)* nível de ensino (técnico, graduação ou básico); *vii)* faixa etária; *viii)* se é natural do mesmo município da instituição ou a outro município.

Fase 2: Pré-processamento. Para a manipulação e preparação dos dados, utilizou-se a linguagem de programação *Python*[2], especificamente a biblioteca *Pandas*[3]. Por meio desta biblioteca, foi possível a remoção de duplicidades, correção de termos, preenchimento de valores nulos, e principalmente, o mapeamento de variáveis categóricas textuais em variáveis categóricas numéricas. Por exemplo, a variável sexo, originalmente armazenada na base como valores 'M' e 'F', foram trocadas por 0 e 1 respectivamente. Este procedimento foi feito para todas as oito variáveis.

5.2. Geração e Avaliação dos Modelos

Os classificadores foram escolhidos com base no levantamento feito na literatura. [Bruce and Bruce 2019] elenca as principais técnicas de classificação existentes na literatura a saber: *Nayve Bayes* (NB), *Decision Tree* (DT), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e algoritmos de *boosting* que consistem da composição de outros modelos, e.g., *gradient boosting*.

Fase 3: Treinamento. Para treino e geração dos modelos, foi definida uma proporção de 70% dos registros de cada campus para treino e 30% para testes e validação. Ao gerar e testar o modelo, uma matriz de confusão 2×2 é criada conforme a Tabela 1. Na matriz de confusão, as células da diagonal principal exibem respectivamente a quantidade de alunos não evadidos preditas (V_n) e evadidos (V_p) preditos corretamente. As células (V_p) e (V_n) somadas representam o quantitativo de acertos do modelo. Em contrapartida, a diagonal secundária exibe os erros de predição dos modelos, no qual (F_p) indica o quantitativos de não-evadidos que o sistema previu como evadido, e (F_n) indica o quantitativo de alunos evadidos que o sistema previu como não-evadido.

Tabela 1. Matriz de confusão usada na avaliação dos modelos.

	Não evasão	**Evasão**
Não evasão	Verdadeiro Negativo (V_n)	Falso Positivo (F_p)
Evasão	Falso Negativo (F_n)	Verdadeiro Positivo (V_p)

Fase 4: Teste e Avaliação. A partir dos valores das matrizes de confusão resultantes de cada modelo criado, é possível calcular as métricas utilizadas nesse trabalho para

[2]https://www.python.org/
[3]https://pandas.pydata.org/

avaliação dos resultados: acurácia (Eq. 1), que representa o percentual de acertos em relação ao total; precisão (Eq. 2), que é a capacidade de prever corretamente os alunos identificados como evadidos; revocação (Eq. 3), que mede a capacidade do modelo de identificar corretamente alunos evadidos em relação aos demais da amostra; e especificidade (Eq. 4), que mede a capacidade do modelo de prever um aluno não-evadido. Como forma de obter uma medida balanceada entre precisão e revocação, também será calculada a medida-F (Eq. 5), que consiste de uma média ponderada entre estas duas métricas, uma vez que podem existir situações em que o modelo produz alta precisão, mas baixa revocação, ou vice-versa.

$$A = \frac{V_p + V_n}{V_n + F_n + V_p + F_p} \quad (1) \qquad P = \frac{V_p}{V_p + F_p} \quad (2) \qquad R = \frac{V_p}{V_p + F_n} \quad (3)$$

$$E = \frac{V_n}{V_n + F_n} \quad (4) \qquad F = \frac{2 \times (P \times R)}{(P + R)} \quad (5)$$

Será utilizado, para melhor compreensão dos resultados, uma visualização gráfica das métricas de desempenho de um modelo através da curva ROC – *Receiver operating characteristic* [Bruce and Bruce 2019] . A curva ROC é uma curva de probabilidade construída em um gráfico de relação entre a taxa de verdadeiros positivos (TPR – *True Positive Rate*) e a taxa de Falsos Positivos (FPR – *False Positive Rate*). TPR tem o mesmo valor que a revocação (Eq. 3), enquanto FPR é dada $1 - E$, i.e., o complemento da especificidade (Eq. 4). Toda a área abaixo da curva ROC, denominada AUC (*Area Under Curve*), representa uma métrica de separabilidade, no qual, quanto mais próximo o valor a 1, maior a capacidade do modelo de predizer amostras em suas respectivas classes.

A metodologia para geração de modelos e cálculos das métricas de avaliação apresentadas aqui, seguiram o conceito de validação cruzada (*cross-validation*), no qual foram executados cinco vezes o procedimento da Figura 1. Em cada uma das 5 execuções os dados foram tomados aleatoriamente, tendo como resultado as médias aritméticas de cada uma das métricas adotadas.

Outra contribuição deste trabalho é a definição do nível de importância das variáveis para os classificadores considerando os dados disponíveis. Tal computação foi realizada após a construção dos resultados finais das métricas anteriormente definidas. O nível da importância torna-se interessante, uma vez que pode sugerir a gestão da instituição qual fator mais pode estar afetando o resultado das probabilidades das predições das evasões feitas. Permitindo, portanto, que o setor pedagógico oriente os demais setores institucionais sobre como atacar esse fator.

6. Resultados e Discussão

Os cinco campi com maior taxa de evasão da rede do IFCE, considerando os dados de 2015 a 2019, são o campus Iguatu com 47,83%, campus Acopiara com 46,20%, Tianguá com 44,66% e, por fim, os campi Acaraú e Quixadá, ambos com 43,73%. Todos os cinco campi possuem taxas acima de 40% de evasão. Estes números são considerados bastante altos e reforçam a importância de haver iniciativas de combate a evasão por parte da instituição.

Na figura 2, é apresentado o gráfico da curva ROC (Seção 5.2) dos melhores classificadores para cada um dos 5 campi. Os classificadores estão destacados na legenda,

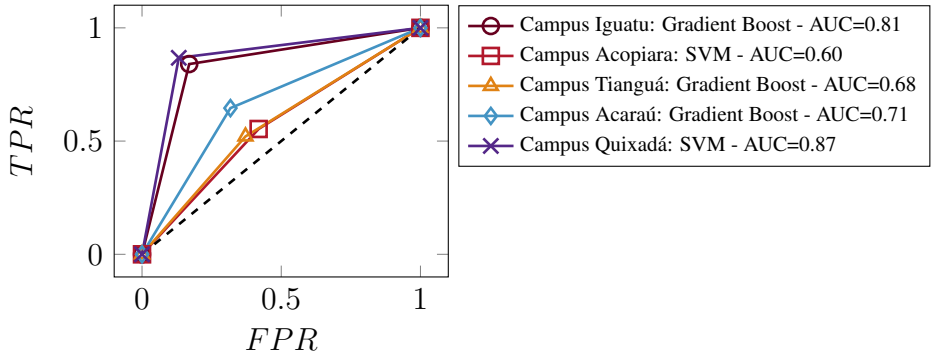


Figura 2. Curva ROC e valores AUC para o melhor classificador de cada campus.

juntamente, com o valor AUC produzido. Perceba que o campus Quixadá obteve a maior área sob a curva ROC, o que significa, em tese, maiores chances de acerto ao designar o potencial de evasão de um determinado aluno. Já para o campus Acopiara, a área sob a curva ROC ficou em torno de 0.55, tendo sido este o pior resultado. Isto pode ser explicado pelo baixo número de estudantes – menor amostragem para treino e teste – em relação a campi como Quixadá e Acopiara, uma vez que estes dois últimos possuem praticamente o dobro de alunos no período.

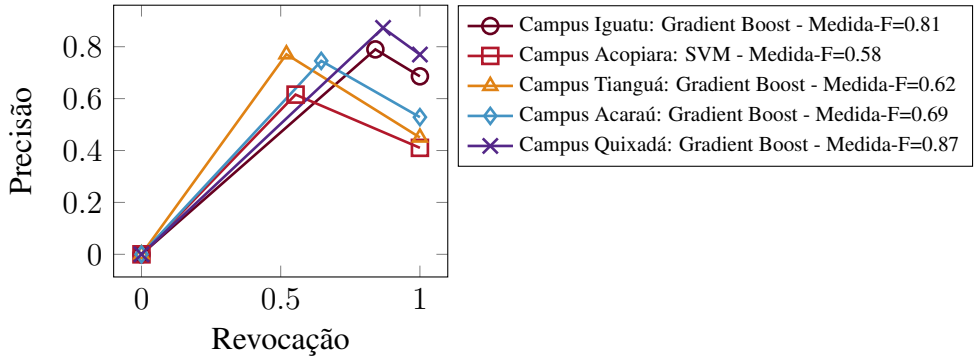


Figura 3. Valores da medida-F para os melhores resultados para cada campus.

O gráfico da Figura 3 exibe, para cada um dos cinco campi, as curvas de precisão vs. revocação, bem como os respectivos melhores índices de medida-F, juntamente com o classificador que os conseguiu gerar (informados na legenda). É possível observar que foram atingidos bons níveis para os campus Iguatu e Quixadá (entre 0.8 e 0.9), níveis razoáveis para Tiaguá e Acaraú (entre 0.6 e 0.7). Para Tianguá, o nível não foi considerado interessante o que pode sugerir a necessidade de mais dados ou até mesmo novas variáveis para o campus em questão.

A Figura 4 mostra o gráfico do nível de importância das variáveis por campus. Por este gráfico, percebe-se que o fato do aluno estar retido impacta de maneira considerável no cálculo da predição de evasão para o campus de Quixadá. Já para o campus Tianguá e Acaraú, a variável nível de ensino foi a que mais se mostrou importante para os respectivos modelos daqueles campi, sugerindo que alunos que ingressam em um determinado nível de ensino têm maior probabilidade de evasão. Para o campus Acopiara, a faixa etária e a etnia foram as variáveis com maior peso. Para o campus Iguatu, as

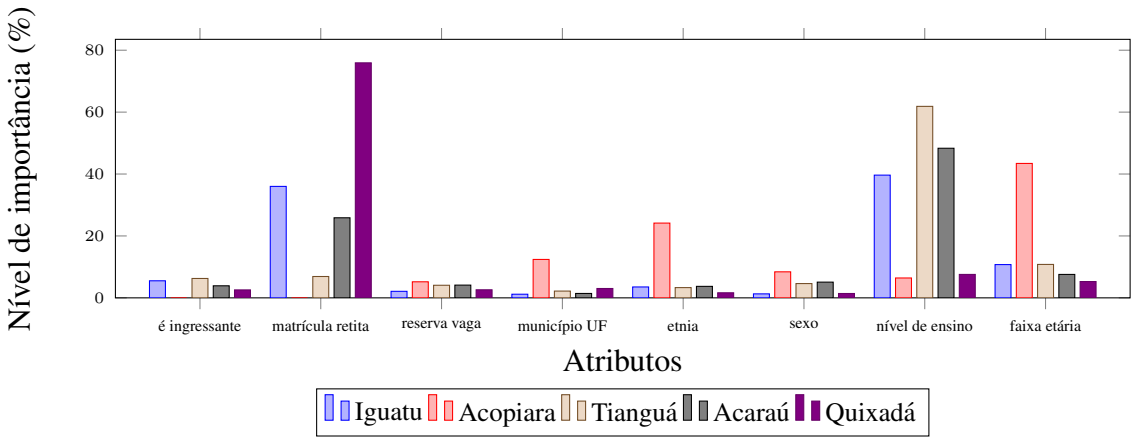


Figura 4. Nível de importância das variáveis por campus.

variáveis que determinam retenção escolar e nível de ensino respectivamente, foram as que mais impactaram na definição dos modelos praticamente com a mesma importância. As variáveis categóricas "são ingressantes?" (ou seja, o discente é ingressante no período ou não) e "Município, UF" (i.e., se o aluno é natural do mesmo município do campus ou não) não tiveram impacto na construção dos modelos de probabilidade.

7. Conclusão

O presente trabalho traçou uma abordagem para a construção de modelos de predição para evasão discente a partir dos dados dos cinco campi com maiores taxas de evasão do IFCE. Os dados foram obtidos a partir do portal IFCE em Números. Foram utilizados alguns dos principais classificadores existentes na literatura para construção e análise dos modelos de predição, tendo os mesmos sido avaliados com as principais métricas atualmente utilizadas pela comunidade de aprendizagem de máquina. Também foi investigado a importância das variáveis (*features*) para os modelos construídos em cada campi, sugerindo assim alguns fatores e características que podem estar influenciando nas respectivas taxas de evasão e melhor nortear as decisões do setor pedagógico e outros setores da instituição.

Como trabalhos futuros, pretende-se criar um sistema Web de apoio à tomada de decisões contendo módulos para extração e limpeza automática dos dados, criação de modelos e interface amigável aos gestores, exibindo análises relevantes, similares às feitas neste trabalho, porém de forma automatizada. A ideia é que cada campus da rede tenha acessos aos dados e análises que lhe dizem respeito acerca da evasão discente, tendo também a opção de agrupar as análises por curso. Isso irá auxiliar de forma positiva a tomada de decisões pedagógicas e administrativas. Além disso, pretende-se obter e considerar variáveis relativas ao rendimento acadêmico dos alunos na geração dos modelos de classificação.

Referências

Amorim, M. J. V., Barone, D., and Mansur, A. F. U. (2008). Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. *XIX Simpósio Brasileiro de Informática na Educação (SBIE)*, 1(1):666–674.

BRASIL (2014). *Documento orientador para a superação da evasão e retenção na Rede Federal de Educação Profissional, Científica e Tecnológica*. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica.

Bruce, P. and Bruce, A. (2019). *Estatistica Pratica Para Cientistas de Dados - 50 Conceitos Essenciais*. Alta Books.

Chung, J. Y. and Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96:346–353.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd edition.

IFCE (2017). Plano estratégico para permanência e êxitos dos estudantes do IFCE. Disponível em: `https://gestaoproen.ifce.edu.br/attachments/download/3052/`. Acesso em 23 de jun. de 2020.

Pró-reitoria do IFCE (2017). IFCE em números. Disponível em: `http://ifceemnumeros.iea.edu.br/`. Acesso em 25 de jun de. 2020.

Queiroga, E., Cechinel, C., and Araújo, R. (2017). Predição de estudantes com risco de evasão em cursos técnicos a distância. *Simpósio Brasileiro de Informática na Educação (CBIE)*, pages 1547–1556.

Rätsch, G. (2004). A brief introduction into machine learning. *Friedrich Miescher Laboratory of the Max Planck Society*.

Rigo, S. J., Cambruzzi, W., Barbosa, J. L. V., and Cazella, S. C. (2014). Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação (RBIE)*, 22(1):132–146.

Sansone, D. (2019). Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, 81(2):456–485.

Saraiva, D., Pereira, S., Gallindo, E., Braga, R., and Oliveira, C. (2019). Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. In *Anais do XXVII Workshop sobre Educação em Computação*, pages 319–333. SBC.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Silva Filho, R. B. and Araújo, R. M. L. (2017). Evasão e abandono escolar na educação básica no brasil: fatores,causas e possíveis consequências. *Educação por Escrito*, 8(1):35–48.

Silveira, G. and Bullock, B. (2017). *Machine Learning: Introdução à classificação*. Casa do Código.

Tan, M. and Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning*, 10(1):11–17.