INTRODUÇÃO

A informática vem se mostrando ao longo do tempo uma importante ferramenta na resolução de problemas em diversas áreas do conhecimento. Tais áreas perpassam por assuntos relacionados à biologia (através da biotecnologia e bioinformática), economia (com aplicações visando, análise de empresas, ganho de capital no mercado financeiro e controle de endividamento e gastos), administração (sistemas de informação e apoio a decisão), área da saúde (e.g., telemedicina, prontuários eletrônicos), dentre outras.

Com relação ao contexto educacional, a evolução da área computacional tornou possível a criação de soluções visando o ensino de componentes curriculares e a educação a distância (EaD), como o desenvolvimento do AVA (Ambiente Virtual de Aprendizagem) tais como o Moodle[1] e o Edmodo[2], além de diversos objetos de aprendizagem como vídeos (com modernas ferramentas de edição), *softwares* educacionais, jogos, simuladores, tutoriais virtuais, *quiz online*, dentre outros (Silva, 2017).

No entanto, o auxílio da informática não se limita apenas às questões do processo de ensino e aprendizagem. Diversas ferramentas e algoritmos vêm surgindo nos últimos anos voltados para uso em áreas como acessibilidade, gestão educacional, e apoio a tomada de decisões pedagógicas.

No que concerne à gestão educacional, a evasão consiste de um dos principais problemas enfrentados em instituições educacionais, não somente no Brasil. Trata-se de um fenômeno global e que pode ser desencadeado por diversos motivos. Esforços a nível governamental e institucional são gastos no sentido de minimizar os índices de evasão. Como parte desse esforço, algumas instituições desenvolveram um documento conhecido como Plano de Permanência e Êxito (PPE), com o objetivo de elucidar informações acerca dos níveis de evasão da instituição, além de elencar as principais causas e possíveis diagnósticos e soluções para promover a permanência do aluno e que este torne-se um egresso de êxito (IFCE, 2017).

As principais causas da evasão podem ser categorizadas em internas a instituição (relacionadas a condições e infraestrutura da instituição), externas (relacionadas a fatores externos como a conjuntura econômica atual), ou causas individuais (relativas a questões individuais, como dificuldades de adaptação ao nível de ensino ou a não-identificação com o curso).

Neste contexto, ferramentas computacionais vêm sendo utilizada no combate à evasão discente. Muitas dessas ferramentas visam facilitar assimilação de determinados conteúdos e componentes curriculares, atuando diretamente no processo de ensino e aprendizagem. (Brasil, 2013) cita algumas iniciativas existentes. Outras ferramentas auxiliam equipes de gestão pedagógica e administrativa a minimizarem os impactos de fatores que causam a evasão.

A inteligência artificial (mais especificamente o aprendizado de máquina) vem se destacando por propor soluções para as mais diversas áreas, como por exemplo a educação, por meio de algoritmos que têm como objetivo principal a predição de discentes com potencial considerável de evasão da instituição. Essa predição é possível devido aos dados pessoais do discente oriundos de seu relacionamento com a instituição.

Aprendizagem de máquina já vem sendo utilizada por diversas corporações públicas e privadas dos mais diversos nichos e setores de serviço. Como exemplo, podemos citar aplicações capazes de identificar com baixo percentual de erro se um email é spam ou não (Dada et al., 2019), ou ainda se um cliente está tendencioso a comprar um certo produto ou não, e até mesmo auxiliar no diagnóstico de uma determinada doença (Silveira e Bullock, 2017).

Com relação a uso de técnicas de aprendizagem de máquina aplicado ao problema da evasão, já existem alguns trabalhos na literatura que versam sobre essa

[1] https://moodle.org/

[2] https://www.edmodo.com/

temática. Como exemplos desses trabalhos, citam-se (Santos et al., 2014) que utilizou algoritmos classificadores baseados em árvore de decisão (*Decision Tree*) para criar modelos para predição precoce de discentes tendenciosos a evasão de cursos de graduação a distância, como também a de alunos com predisposição a reprovarem disciplinas. O trabalho desenvolvido em (Kantorski et al., 2016) foca na construção de modelos voltados para a evasão discente em instituições públicas de ensino. Para isso, o trabalho considerou dados oriundos de discentes dos cursos de Zootecnia e administração de uma universidade pública, bem como a utilização de classificadores baseados em árvore de decisão e algoritmos probabilísticos de Bayes (*Naive Bayes*) para a construção e validação dos modelos. Ambos os trabalhos obtiveram resultados expressivos em relação às respectivas realidades. Em âmbito internacional, (Sansone, 2019) combina teorias econométricas com algoritmos de aprendizagem de máquina para construção de modelos de predição para evasão do nono ano (9^{th} *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*).

O presente trabalho visa utilizar as técnicas de aprendizagem de máquina para analisar dados oriundos do Instituto Federal do Ceará (IFCE), tendo como estudo de casos os campi de Paracuru e Jaguaruana, os quais estão entre os que detém os maiores índices de evasão da rede. A partir de dados previamente coletados, pretende-se utilizar algoritmos de classificação mais amplamente aceitos pela comunidade para gerar modelos que possibilitem predizer a evasão dos discentes no campi. Além disso, são analisados e discutidos os resultados, apontando possíveis melhorias e eventuais trabalhos futuros.

MATERIAL E MÉTODOS

Para atingir aos objetivos propostos, primeiramente foi feito um levantamento do estado da arte em duas áreas de estudo principais: área educacional, focando na evasão; área computacional, focando no problema de classificação, bem como conceitos, metodologias, técnicas e algoritmos relacionados.

Baseado nesse levantamento prévio, o seguinte fluxo de trabalho foi definido (Figura 1):

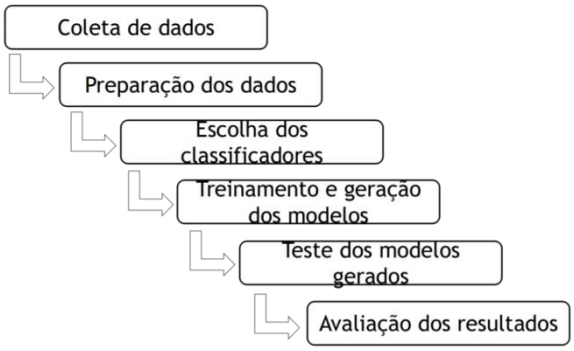

Figura 1. Etapas do fluxo de trabalho

Coleta de dados. Esta etapa consistiu da extração de dados públicos dos alunos dos campi Paracuru e Jaguaruana do Instituto Federal do Ceará (IFCE), a partir do acesso a plataforma IFCE em números (Pró-reitoria IFCE, 2017). Todos os dados foram obtidos em formato tabular de planilha eletrônica. 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.

Preparação dos dados. Para a manipulação e preparação dos dados, utilizou-se a linguagem de programação *Python*, 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.

Escolha dos classificadores. Os classificadores foram escolhidos com base no levantamento feito na literatura. Segundo (Şara et al., 2015), algoritmos como *SVM* e *RandomForest* têm obtido bons resultados em problemas de classificação. Além destes, optou-se por incluir os clássicos algoritmos Bayesiano (*NaiveBayes*), *KNN* (*K-Nearest-Neighbors*) e o algoritmo de gradiente (*Gradient Boosting*). Todos estes classificadores foram acessados a partir da biblioteca *scikit-learn*[4], que consiste de um conjunto de módulos e funcionalidades voltados para soluções em aprendizagem de máquina.

Treinamento e geração dos modelos. Para cada um dos dois campi, os dados foram separados em uma proporção de 70% para o treinamento e geração dos modelos, e 30% para os testes e validação dos modelos gerados por cada classificador. A Tabela 1 mostra os

[3] https://pandas.pydata.org/

[4] https://scikit-learn.org/

quantitativos e categorias dos dados dos alunos de cada um dos dois campi.

Tabela 1. Quantitativo de dados por campi.

	Campus Paracuru	Campus Jaguaruana
Matriculados	643	435
Retidos	51	231
Concluídos	570	347
Evadidos	678	614
Total	1942	1627

Ao analisar a Tabela 1, percebe-se que, em ambos os campi, o número de não evadidos (matriculados, retidos e concluídos) é bem maior que o número de evadidos, chegando este último a ter um terço do valor da classe oposta. Neste caso, é possível afirmar que os dados estão desbalanceados. A partir de bases de treino desbalanceadas, classificadores podem gerar modelos tendenciosos (com viés) favorável a uma das classes.

A fim de evitar o problema do enviesamento, foram considerados, para ambos os campi, apenas os alunos com os status de "concluído" e de "evadido", respectivamente, observando uma quantidade de dados relativamente similar entre essas classes e garantindo assim o balanceamento dos dados de treinos.

Teste dos modelos gerados. Na fase de teste e avaliação, os resultados obtidos dos modelos gerados pelos classificadores foram analisados, ressaltando-se erros e acertos de cada para cada instância de teste. Desta análise, foram construídas as respectivas matrizes de confusão. A Figura 2 exibe o formato geral desse tipo de matriz. Via de regra, a diagonal principal (destacada em azul) exibe os acertos do modelo, enquanto que a diagonal invertida (destacada em vermelho) exibe os erros.

Os chamados verdadeiros negativos representam os acertos do modelo com relação aos discentes que não evadiram. Os verdadeiros positivos são os acertos do modelo com relação aos discentes que evadiram. Em contrapartida, os falsos positivos consiste da quantidade de discentes que o modelo acusou serem evadidos, mas que na verdade, não o são. Os falsos negativos por sua vez são os discentes que se evadiram, e o modelo acusou que o mesmo não o fizeram.

	Não-Evasão	**Evasão**
Não-Evasão	Verdadeiros negativos (V_n)	Falsos positivos (F_p)
Evasão	Falsos negativos (F_n)	Verdadeiros positivos (V_p)

Figura 2. Matriz de confusão.

A partir da matriz de confusão, foram extraídos 5 medidas comumente utilizadas para avaliação dos resultados de um modelo de classificação. São estas a acurácia, a precisão (ou exatidão), a revocação (ou sensibilidade), a especificidade, e a medida-*F* (ou score F1). As fórmulas (1), (2), (3), (4) e (5) representam respectivamente a forma de calcular essas medidas.

$$A = \frac{V_p + V_n}{V_n + F_n + V_p + F_p} \quad (1)$$

$$P = \frac{V_p}{V_p + F_p} \quad (2)$$

$$R = \frac{V_p}{V_p + F_n} \quad (3)$$

$$E = \frac{V_n}{V_n + F_n} \quad (4)$$

$$F = \frac{2 \times (P \times R)}{(P + R)} \quad (5)$$

Avaliação dos resultados. A acurácia representa a porcentagem de acerto total do modelo. Esta métrica foi utilizada durante um considerável tempo como sendo a única medida de avaliação da qualidade dos resultados de um modelo de classificação. Entretanto, (Bruce e Bruce, 2019) mostra que tal medida por si só pode mascarar algumas falhas e fraquezas do modelo.

Para exemplificar o problema de utilizar apenas a acurácia como métrica de qualidade, basta imaginar uma situação em que há uma tendência de maioria da amostra para uma das duas classificações. Suponha uma situação de criação de um modelo para tentar predizer fraudes em cartões de crédito. Neste caso, o número de casos considerados como fraude pode ser bem pequeno em relação ao número de casos considerados regulares. Logo, utilizando um modelo que, para toda amostra de cartão, sempre responda como sendo um cartão regular, tende a ter um nível alto de acurácia. No entanto, tal modelo certamente não é interessante para resolução do problema. A Tabela 2 mostra o valor de 99,33% de acurácia para um exemplo de um modelo que sempre responde como cartão regular para qualquer amostra, em que o

número de cartões regulares é de 300.000 (trezentos mil) e o de cartões ilegais é de apenas 2000 (dois mil). A partir de cenários como estes, foram adotadas as medidas de precisão, revocação, especificidade, e a medida-*F*.

No contexto do problema de classificação entre evasão e não evasão, a precisão representa a capacidade do modelo de prever corretamente os alunos identificados como evadidos. A especificidade mede a capacidade do modelo de prever um aluno não-evadido. Já a revocação mede a capacidade do modelo de identificar corretamente alunos evadidos em relação aos demais da amostra.

Tabela 2. Exemplo em que a acurácia não é boa métrica de qualidade: todos os cartões classificados como regulares e acurácia muito alta.

Cartões regulares	**Cartões ilegais**	**Tendência de acurácia resultante**
300.000	2.000	300.000 / 302.000 = 0,9933

A medida-*F* representa uma média harmônica entre os níveis de precisão e revocação do modelo. Esta medida se torna interessante uma vez que existe um certo "*trade-off*" entre essas medidas. Quando um modelo considera um baixo espaço de busca dentre o disponível (classificar menos amostras), tende a acertar mais predições (ter uma precisão maior). A medida em que o modelo é alterado e passa a considerar mais elementos da amostra, tende a ter uma maior probabilidade maior de errar na predição, diminuindo assim a precisão. A medida-*F* consiste de um esforço em quantificar esse "tradeoff" em uma só métrica.

Considerando o estudo da literatura na área (Bruce e Bruce, 2019) (Gron, 2019), as métricas supracitadas, bem como gráficos posteriormente construídos, foram utilizadas para embasar as análises feitas no presente trabalho.

RESULTADOS E DISCUSSÃO

A Tabela 3 apresenta os valores obtidos de acurácia (A), precisão (P), revocação (R), especificidade (E) e medida-*F* (F) para o campus Paracuru. Cada valor foi obtido da média aritmética entre os valores de cada uma das 5 (cinco) iterações da validação cruzada. Os classificadores estão representados da seguinte forma: DT - *Decision Tree* (árvore de decisão), NB - *Nayve Bayes*, KNN - *K-Nearest-Neighbors*, SVM - *Support Vector Machin*e, GB - *Gradient Boosting*.

A Tabela 4 apresenta os valores obtidos para o campus Jaguaruana a partir das mesmas métricas aplicadas ao campus Paracuru.

A partir dos dados obtidos, é possível afirmar que o classificador GB (*Gradient Boosting*) obteve os melhores resultados, tanto no campus Paracuru como no campus Jaguaruana. O princípio deste classificador é executar vários "modelos de aprendizagem" (modelos fracos) de forma iterativa, combinando-os, de maneira sequencial, e gerando um único modelo de aprendizagem considerado melhor (forte). (Mayer, 2014) mostra que algoritmos de boosting podem de fato serem promissores em relação a algoritmos tradicionais. Por tal procedimento, algoritmos de boosting também são conhecidos como algoritmos de agrupamento de modelos. Geralmente tais modelos são baseados em árvores de decisão.

Tabela 3. Quantitativo de dados para o campus Paracuru.

Classificador	**A**	**P**	**R**	**E**	**F**
DT	0.66	0.70	0.64	0.61	0.67
NB	0.66	0.72	0.60	0.61	0.65
KNN	0.66	0.69	0.69	0.63	0.69
SVM	0.64	0.67	0.65	0.60	0.66
GB	0.68	0.71	0.71	0.65	0.71

Tabela 4. Quantitativo de dados para o campus Jaguaruana.

Classificador	**A**	**P**	**R**	**E**	**F**
DT	0.75	0.81	0.79	0.65	0.80
NB	0.69	0.83	0.65	0.56	0.73
KNN	0.76	0.79	0.85	0.70	0.82
SVM	0.75	0.77	0.85	0.71	0.80
GB	0.78	0.81	0.86	0.72	0.83

A Figura 3 exibe graficamente as curvas de tradeoff entre precisão e revocação para os resultados obtidos a partir do classificador de *Gradient Boosting*.

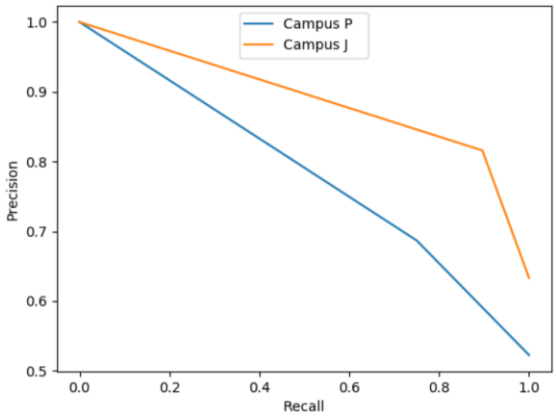

Figura 3. Precisão x Revocação - Campi Paracuru e Jaguaruana.

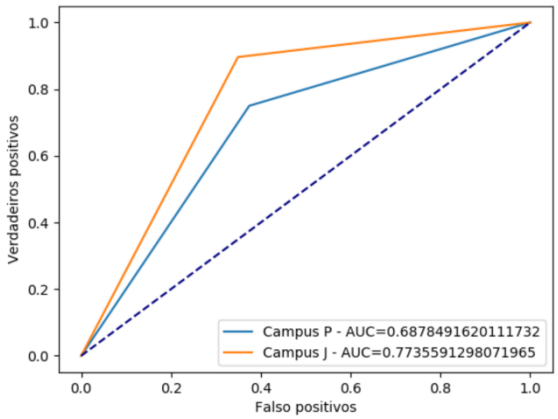

Figura 4. Curva AUC – Campus Paracuru e campus Jaguaruana.

Percebe-se que a Tabela 4 apresenta resultados melhores que a Tabela 3 e, também, visualmente pelo gráfico da Figura 3, que os níveis dos indicadores do campus Jaguaruana foram levemente melhores do que o campus Paracuru. A provável explicação para isso é uma menor diversificação no perfil do aluno de cada campus. Analisando os dados do campus Paracuru, percebe-se que há uma maior variedade de dados entre seus discentes do que no campus Jaguaruana. Cita-se como exemplo que o campus Jaguaruana não possui cursos de graduação, diferentemente do campus Paracuru que possui, além de cursos técnicos e básico, dois cursos de graduação. Além disso, o campus Paracuru também possui uma maior variedade de categorias de cotistas cadastradas em relação ao campus Jaguaruana.

Um conceito que vem sendo utilizado na visualização de métricas de desempenho de um modelo é a chamada curva ROC – Receiver operating characteristic (Bruce e Bruce, 2019). A curva ROC é uma curva de probabilidade construída em um gráfico de relação entre os valores de verdadeiros positivos e os falsos positivos. Toda a área abaixo da curva ROC, frequentemente referenciada como 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 Figura 4 exibe um gráfico contendo a curva ROC para os modelos gerados pelo *Gradient Boosting* nos dois campi analisados, bem como seus respectivos valores de AUC. O mesmo gráfico mostra valores de verdadeiros positivos (eixo das ordenadas) e os falsos positivos (eixo das abscissas). Verifica-se um maior valor de AUC para o campus Jaguaruana, significando que, em tese, os modelos gerados para o campus Jaguaruana são mais assertivos e tem uma maior capacidade de separar os prováveis alunos evadidos dos não-evadidos, em relação aos gerados para o campus Paracuru.

CONCLUSÃO

O presente trabalho traçou o histórico para a construção de modelos de predição dos níveis de evasão a partir dos dados de dois campi do IFCE. Os dois campi analisados estão entre os maiores índices de evasão de toda a rede o que motivou a escolha dos mesmos.

Os modelos foram treinados e testados utilizando dados extraídos do portal IFCE em números, e avaliados através das métricas frequentemente utilizadas na literatura.

Os modelos gerados a partir do campus Jaguaruana foram levemente mais bem avaliados do que os de Paracuru, e isso provavelmente se aplica ao perfil dos alunos e até mesmo do próprio campus ser menos diverso, com menos cursos e níveis ofertados.

Uma possibilidade de trabalho futuro seria considerar novas variáveis que podem ter alguma relação com o fenômeno da evasão, a exemplo dos dados de rendimento acadêmico dos discentes. Além disso, pode-se investigar maneiras de melhorar os resultados dos classificadores, através de possíveis e novos valores de parametrizações. Por fim, a construção de um sistema computacional completo (incluindo o desenvolvimento de interfaces gráficas) que poderá ser utilizado pela gestão da instituição no auxílio à tomada de decisões.

AGRADECIMENTOS

Os agradecimentos vão para o Instituto Federal do Ceará, em especial aos diretores do campi Paracuru e Jaguaruana pelo apoio dado.

REFERÊNCIAS

Brasil (2013). Guia de tecnologias educacionais da educação integral e integrada e a articulação da escola com seu território. BRASIL. Ministério da Educação.

Bruce, P.; Bruce, A. Estatística prática para cientista de dados - 50 conceitos essenciais. (Juliana de

Oliveira, Thiê Alves e Illysabelle Trajano). Alta Books, Rio de Janeiro, 2019.

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., Aji-buwa, O. E. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), 2019.

Gron, A. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd. ed., (Nicole Tache) O'Reilly Media, Inc, 2019.

IFCE. Plano estratégico para permanência e êxito dos estudantes do IFCE. IFCE. Instituto Federal de Educação, Ciência e Tecnologia do Ceará. Pró-reitoria de Ensino - PROEN, Fortaleza, 2017.

Kantorski, G. Z., Flores, E. G., Hoffmann, I. L., Schmitt, J. A., Barbosa, F. P. Predição da evasão em cursos de graduação em instituições públicas. Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), 27, p. 906-915, 2016.

Mayr, A. Binder, H., Gefeller, O., Schmid, M. The evolution of boosting algorithms. Methods of information in medicine, v. 53, n. 06, p. 419-427, 2014.

Pró-reitoria IFCE. IFCE em números. http://ifceemnumeros.ifce.edu.br/. Acesso em 25 jun, 2019.

Şara, N. B., Halland, R., Igel, C., Alstrup, S. High-school dropout prediction using machine learning: A Danish large-scale study. In ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence, 1, p. 319-324, 2015.

Sansone, D. Beyond early warning indicators: high school dropout and machine learning. Oxford Bulletin of Economics and Statistics, 81, p. 456-485, 2019.

Santos, R. N.; Siebra, C. de A.; Soares, Oliveira, E. D. S. Uma Abordagem Genérica de Identificação Precoce de Estudantes com Risco de Evasão em um AVA utilizando Técnicas de Mineração de Dados. XIX Congresso Internacional de Informática Educativa, 10, p. 794-799, 2014.

Silva, R. S. Gestão de EAD - Educação a distância na era digital . (Rubens Prates). Novatec, São Paulo, 2017.

Silveira, G. Bullock, B. Machine Learning -introdução a classificação. (Adriano Almentida e Vivian Matsui). Casa do código, São Paulo, 2017.