

## QUEM SERÁ O PRÓXIMO ALUNO A ABANDONAR O CURSO?

Carla Ponciano <sup>1</sup>; Thiago Rezende <sup>2</sup>; Jânio da Silva <sup>3</sup>; Luciana Silva <sup>4</sup>

1 Bolsista (CNPq), Sistemas de Informação, IFMG, Campus Ouro Branco; si.carla.ponciano@outlook.com

2 Sistemas de Informação, IFMG, Campus Ouro Branco

3 Pesquisador do IFMG, Campus Ouro Branco

4 Orientador: Pesquisador do IFMG, Campus Ouro Branco; luciana.lourdes.silva@ifmg.edu.br

### RESUMO

Atualmente, as instituições de ensino no Brasil têm enfrentado um alto índice de evasão de alunos. No ensino superior, a taxa de evasão é de aproximadamente 22%. A evasão gera prejuízos significativos para as instituições de ensino tanto particulares quanto públicas. Além disso, o próprio aluno é afetado ao deixar o curso em andamento, comprometendo seu futuro profissional. Neste projeto, o objetivo é desenvolver uma ferramenta baseada no histórico de alunos matriculados e prever quais alunos estão prestes a abandonar o curso. Desta maneira, a instituição pode diagnosticar o motivo e auxiliar o aluno, evitando assim a desistência no curso. Os modelos gerados pelo Multilayer Perceptron, Random Forest (RF) e AdaBoost apresentaram 94% de acurácia. RF foi o único modelo a obter 94% de F-Score e Kappa maior que 0.6.

**Palavras-chave:** Alunos. Evasão. Predição. Disciplinas. Desistência. Aprendizado de Máquina.

### INTRODUÇÃO:

Um dos grandes problemas enfrentados pelas instituições brasileiras hoje é o alto índice de evasão dos seus alunos. Segundo Lobo (2017), o ensino superior no Brasil tem aproximadamente uma taxa de 22% de evasão nas instituições privadas e um pouco menos nas instituições públicas. Esse alto índice de evasão pode afetar não só as instituições de ensino particulares que têm os seus orçamentos duramente afetados devido a baixa quantidade de mensalidades pagas, mas também as federais que tem como objetivo, entre outros, formar mão de obra qualificada para o mercado. Por exemplo, no período de 2013 a 2018, em um campus do Instituto Federal de Minas Gerais registrou um total de 1.500 alunos, onde 16,39% se evadiram dos seus respectivos cursos. Outro grande afetado por este problema é o próprio aluno que ao deixar o curso compromete o seu futuro profissional, devido à carência do mercado por profissionais qualificados em áreas como Sistemas de Informação.

Existem inúmeros motivos que levam um aluno a abandonar o curso superior, como a falta de recursos financeiros (transporte, alimentação e moradia), conflito de horários com o emprego, problemas de saúde, dificuldades de aprendizado, desmotivação, etc. A percepção precoce da possível tendência de desistência deste aluno por parte da instituição talvez possa evitar a sua evasão, auxiliando-o a contornar sua situação. No entanto, detectar este cenário em um universo de milhares de alunos não é uma tarefa fácil.

### Objetivo Geral

Neste projeto será desenvolvido um sistema de predição baseado em técnicas de aprendizado de máquinas para prever alunos que tendem a abandonar o curso em que está matriculado. Com esta técnica, responsáveis pelo acompanhamento de alunos (ex: pedagogos) talvez possam auxiliar o aluno na solução do problema e evitar o abandono do curso. Como as razões que levam um aluno a desistir do curso podem ser bem complexas, a técnica tem como objetivo auxiliar na redução do número de evasão.

### TRABALHOS RELACIONADOS

Atualmente, o uso de técnicas de mineração de dados para detectar alunos com alto risco de abandonar o curso em que estão matriculados tem despertado a atenção da comunidade acadêmica, SILVA & ADEODATO (2012); MANHÃES et al. (2014); DIGIAMPIETRI et al. (2016); SOUZA (2008).

SILVA & ADEODATO (2012) usam regressão logística para identificar alunos com alto risco de evasão. Eles conduziram o estudo com base em diferentes cursos da Universidade Federal de Pernambuco. Eles se basearam em várias informações dos alunos, como notas médias dos 1º e 2º semestres, taxa de reprovação de um semestre para o outro e taxa de aprovação nas provas finais.

MANHÃES et al. (2014) usam um modelo baseado em árvores de decisão para monitorar o andamento dos alunos e avaliar o risco de evasão. Eles conduziram o estudo baseado nas notas e frequências de cada disciplina e situação dos alunos em três cursos de engenharia. O modelo ultrapassou 90% na taxa de acerto.

DIGIAMPIETRI et al. (2016) usam o algoritmo Rotation Forest para detectar alunos com alto risco de evasão. Eles extraíram o histórico escolar de 1.896 alunos, como disciplinas cursadas, frequência e notas. Eles observaram que o atributo chave para obter uma classificação mais precisa foi o semestre que o aluno foi aprovado em uma disciplina específica.

THAKER et al. (2020) explora métodos para identificar de maneira automática os materiais recomendados de livro-texto que são mais relevantes e apropriados ao aluno. Mais especificamente, eles avaliaram como incorporar o estado de conhecimento atual do aluno em conceitos de domínio (ex: recuperação de informação, sistema de informação e indexação baseada em conteúdo são exemplos de conceitos de domínio) associados com a atividade para recomendar seções personalizadas para cada aluno. Os resultados mostraram que a combinação com os estados de conhecimento do aluno tende a aumentar significativamente a qualidade das recomendações, em relação a recomendações tradicionais baseadas em conteúdo.

YAO et al. (2020) definiram um modelo para analisar a procrastinação de alunos em MOOCs. Especificamente, eles buscam encontrar associação entre procrastinação dos estudantes e atividades dos estudantes dentro e entre diferentes tipos de materiais online. Para isso, eles modelam a sequência de interação entre cada aluno e cada módulo do curso. Eles reportaram que existem dependências entre atividades históricas dos estudantes e as futuras também quando diferentes tipos de materiais de aprendizado estão envolvidos.

O que difere este projeto dos demais está na maneira de construir o modelo e na base de dados. Os trabalhos existentes usam somente cursos superiores, enquanto nós iremos investigar também o uso de informações dos alunos matriculados nos cursos técnicos integrados.

## **METODOLOGIA:**

Como a proposta é elaborar um modelo de predição que auxilie no combate à evasão escolar, o primeiro passo é definir quais serão as informações adotadas para gerar o preditor. Em resumo, esta pesquisa é feita em quatro etapas:

- *Coleta dos Dados* - a base de dados consiste no histórico acadêmico dos alunos que se matricularam em todas as modalidades durante o período de **2013 a 2018** no campus onde o projeto está sendo conduzido. Para cada curso, a secretaria forneceu um arquivo CSV descrevendo todos os dados funcionais dos alunos. Especificamente, para cada aluno tem-se: matrícula, nome, curso, data de ingresso, data de egresso (se existir), código e nome das disciplinas cursadas, situação do aluno nas disciplinas cursadas, nota, falta, semestre e ano que as cursou. Por fim, a base de dados possui todos os alunos com as situações: matriculado, trancado, concluído, desligado;
- *Pré processamento* - Preparar a base de dados para alimentar os algoritmos de classificação a serem testados;
- *Modelo* - Calibrar os parâmetros dos algoritmos, treinar o modelo de classificação;
- *Ferramenta* - Desenvolver uma ferramenta que permita o usuário avaliar os alunos ativos.

A seguir são apresentados detalhes sobre os passos de pré processamento, modelo e ferramenta.

- **Classes** - Para definir a classe da base de dados que será usada para a geração do modelo foram considerados somente os alunos inativos (os que concluíram o curso e os que desligaram). Assim, tem-se uma base binária, ou seja, cada aluno deve estar rotulado como Formado ou Desligado. Com base nesses rótulos, o algoritmo irá detectar padrões que existem nos atributos dos alunos que se desligaram do curso que se diferenciam dos alunos formados.
- **Derivação de Features (atributos)** - A partir da base foi aplicada a técnica de derivação de features a partir das features primárias. Por exemplo, se um aluno do curso de SI cursou a disciplina POO três vezes. É possível gerar features como `POO_numVezezCursada` (tipo inteiro). Logo, as features deste aluno para POO são: `POO_carga_horaria`, `POO_numVezezCursada`, `POO_nota`, `POO_aulas_ministradas`, `POO_nova_turma`, `POO_faltas`, `POO_recuperacao` (boolean), `POO_ano`, `POO_semestre`.
- **Transformação dos Dados** - Nesta etapa são aplicadas técnicas de normalização (entre [0:1]) e transformação dos dados. Por exemplo, as notas foram classificadas em categorias de A a F. Esta etapa resultou em quatro bases para treinamento e testes após aplicar técnicas de transformação (sem nenhuma aplicação, log, raiz quadrada e potência).
- **Missing Values** - Como existem várias disciplinas que muitos alunos não cursaram, existem muitos atributos que estão sem dados. Para tais atributos foi definido o valor -1, caracterizando assim que a disciplina ainda não foi cursada.
- **Seleção de Features** - Em seguida, foi feita a análise dos atributos considerando CfsSubsetEval, Correlação de Pearson e Análise Principal de Componentes. Para cada configuração de base gerada na etapa anterior é feita a aplicação de técnicas de seleção de features.
- **Algoritmos** - Baseando-se no passado, os algoritmos de aprendizado de máquinas conseguem prever o futuro. Nesta etapa é investigada diferentes estratégias para obter o bom desempenho do modelo de predição. Existem diferentes algoritmos que são baseados em árvores, redes neurais, ensembles e stacking. Neste trabalho os algoritmos estudados são: Multilayer Perceptron, AdaBoost, J48, Random Forest, Naive Bayes, Cost Sensitive Classifier.
- **Validação** - Após gerar o modelo de predição com os ex-alunos (Desligados e Formados) é feita uma validação com os alunos ativos, onde serão aplicados no modelo para detectar quem destes alunos seguem os padrões dos que se desligaram.
- **Tomada de Decisões** - A ferramenta deve apresentar alertas que indicam a probabilidade de determinado estudante desistir do curso. Assim, auxiliar a área pedagógica na antecipação de alguma intervenção. Para esta etapa, serão disparados três tipos de alertas: i) saída iminente do aluno (alerta vermelho); ii) forte indício que o aluno segue uma tendência de abandonar o curso (alerta amarelo); iii) o aluno segue o fluxo normalmente (alerta verde). Esses alertas são calculados baseando-se na saída de predição do modelo que indica a probabilidade de um aluno fazer parte da classe sim ou não (valores [0:1] ). Baseado neste valor, é feita a classificação da seguinte maneira: entre [0:0.4] - alerta verde; entre (0.4:0.7) - alerta amarelo; entre (0.7:1] - alerta vermelho.
- **Relatório** - Além de permitir o usuário de listar os alunos e ordená-los por alerta, a ferramenta fornecerá uma opção de emitir relatório dos alunos a serem avaliados pelo usuário (que pode ser um pedagogo ou mesmo um coordenador de curso) agrupados pelos alertas. Desta maneira, é possível o usuário encaminhar resultados para demais responsáveis.

## Tecnologia

A ferramenta está sendo implementada com tecnologias amplamente adotadas pelo mercado atualmente. A seguir breve descrição das mesmas:

- **Backend** - A arquitetura adotada para o *backend* é micro serviços utilizando a linguagem Java e GraphQL (query language).

- O *Frontend* está sendo desenvolvido em Vue.js.
- *Database* - O sistema de gerenciamento de banco de dados adotado é o PostgreSQL.
- Implementação de virtualização de *containers* - A aplicação consistirá de três *containers Frontend, Backend e Database*. Será usado o Docker Compose para definir e rodá-los.
- Para Integração Contínua foi adotado o Jenkins.

### Métricas de Avaliação

Neste trabalho é adotado o uso de precisão, revocação, F1 e AUC (área sob a curva ROC) para avaliar a qualidade do modelo gerado. A seguir algumas descrições sobre cada métrica:

*Área sob a curva ROC (AUC)*: é uma medida de discriminação. Basicamente, se um modelo gerado possui uma área elevada sob a curva ROC, o resultado sugere que o modelo é capaz de prever com precisão. HOSMER and LEMESHOW (2000) apresentam os seguintes valores AUC:

Valor AUC	Significado
AUC = 0,5	Sem discriminação (ou seja, o mesmo que jogar uma moeda para o alto)
$0,7 \leq AUC < 0,8$	Discriminação aceitável
$0,8 \leq AUC < 0,9$	Discriminação excelente
$AUC \geq 0,9$	Discriminação fora de série (mas, extremamente rara)

*F1 Score*: média harmônica entre precisão e revocação.

*Quantificadores*: Suponha uma base com a classe A e B. Suponha também que deseja-se calcular a precisão e revocação da classe A. Logo, os quantificadores possuem o seguinte significado:

- TP (True Positives): Alunos classificados como pertencentes a classe A e que são realmente da classe A.
- FP (False Positives): Alunos classificados como pertencentes à classe A, mas que pertencem à classe B.
- P (Positives): Número de alunos que pertencem a classe A.

*Precisão*: Em resumo, a precisão é calculada por  $TP / (TP + FP)$

*Revocação*:  $TP/P$ . É calculada para avaliar a cobertura que um determinado modelo obtém sobre um conjunto de predições.

*Kappa (Cohen's Kappa)*: é semelhante a acurácia de classificação, particularmente útil em base de dados desbalanceados (a quantidade de uma classe é significativamente maior que a outra). Intervalos e nível de significância: fraca [0:0.2], razoável [0.21:0.4], moderada [0.41:0.6], forte [0.61:0.8], quase perfeita [0.81:1].

### RESULTADOS E DISCUSSÕES:

**Transformação dos Dados**: Os resultados não se destacaram por adotar uma etapa de transformação dos dados. Logo, os dados adotados foram os naturais e para cada algoritmo foram aplicados a normalização entre [0:1].

**Seleção de Features:** A métrica selecionada foi a CfsSubsetEval, apesar de todas apresentarem resultados semelhantes.

**Algoritmos:** Os Ensembles usados são AdaBoost, Random Forest e Cost Sensitive. Para AdaBoost e Cost Sensitive é possível definir classificadores a serem usados. Para estes, foram considerados todos os algoritmos definidos neste trabalho e foi reportado somente o que obteve melhor resultado.

A Tabela a seguir mostra os resultados obtidos por todos os algoritmos considerados neste trabalho para o curso de Licenciatura em Computação. Na base de dados considerada do curso possui 140 alunos, onde destes, 13 foram desligados. Logo, temos uma base bastante desbalanceada e o baseline para esta base é de 0.91 de acurácia. Em outras palavras, o modelo deve ter uma acurácia maior que 0.91 para ser interessante. A linha Acurácia mostra que somente a Multilayer Perceptron (MP), Random Forest (RF), Naive Bayes (NB) e AdaBoost com o Random Forest (AB) conseguiram resultados acima do baseline. O RF foi o modelo que conseguiu o maior F-Score e Kappa, 0.94 e 0.61 respectivamente. Na linha AUC, podemos observar que somente Naive Bayes e Multilayer Perceptron obtiveram AUC maior que 0,7 (aceitável).

LICENCIATURA EM COMPUTAÇÃO								
Métricas	ZeroR	Multilayer Perceptron	AdaBoost J48	J48	Random Forest	Naive Bayes	Cost Sensitive Classifier - RF	AdaBoost RF
Acurácia	0.91	<b>0.94</b>	0.91	0.91	<b>0.94</b>	<b>0.92</b>	0.91	<b>0.94</b>
Precisão	?	0.93	0.90	?	0.94	0.91	0.91	0.93
Revocação	0.91	0.94	0.91	0.91	0.94	0.92	0.91	0.94
F-Score	?	0.93	0.91	?	<b>0.94</b>	0.91	0.91	0.93
AUC	0.41	<b>0.77</b>	0.66	0.41	0.64	<b>0.79</b>	0.65	0.65
Kappa	0	0.57	0.44	0	<b>0.61</b>	0.44	0.49	0.54

Fazendo uma análise mais detalhada, o objetivo é encontrar os alunos ativos com perfil daqueles que evadiram o curso (que é um número significativamente menor que os que concluíram, no ponto de vista de problemas de classificação). Logo, os modelos que detectam o maior número deste perfil são mais interessantes para a ferramenta. Por exemplo, o RF, MP e Cost Sensitive detectaram 54% dos alunos que evadiram o curso de LC. Já o Naive Bayes e o AdaBoost detectaram 38% dos alunos.

## CONCLUSÕES:

Neste estudo foi apresentado o resultado obtido com a aplicação dos algoritmos na base de dados relacionada aos alunos do curso de Licenciatura em Computação. Considerando que as bases de dados de todos os cursos seguem o mesmo padrão de desbalanceamento das classes, uma abordagem como Over Sampling

possa apresentar resultados superiores aos reportados aqui neste estudo. Os modelos MP, RF e AdaBoost apresentaram 94% de acurácia. RF foi o único modelo a conseguir 94% de F-Score e mais de 0.6% de Kappa.

Os próximos passos podem ser listados da seguinte maneira:

1. Aplicar Informed Over Sampling (SMOTE) para balancear a base de dados e re-aplicar todos os testes e coletar as métricas. Se os resultados se demonstrarem promissores em relação aos apresentados neste artigo, o SMOTE será adotado para os demais cursos;
2. Aplicar os algoritmos nos demais cursos;
3. Descrever as features consideradas no modelo final e avaliar como elas melhoram o resultado do classificador;
4. Integrar a implementação do melhor algoritmo com a ferramenta;
5. Submeter artigo para conferência ou revista.

## REFERÊNCIAS BIBLIOGRÁFICAS:

LOBO, Maria Beatriz de Carvalho Melo. "Esclarecimentos metodológicos sobre os cálculos de evasão". Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia. Mogi das Cruzes, SP: 2011. Disponível em: Acesso em: 29 de outubro de 2019.

SOUZA, Solange Lima de. "Evasão no Ensino Superior: Um Estudo Utilizando a Mineração de Dados como Ferramenta de Gestão do Conhecimento em Um Banco de Dados Referente À Graduação de Engenharia". Dissertação de mestrado em Ciências, COPPE/UFRJ, 2008.

MANHÃES, Laci Mary Barbosa; CRUZ, S.M.S.; ZIMBRÃO, G. "WAVE: An Architecture for Predicting Dropout in Undergraduate Courses Using EDM". 29th ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING (SAC). ACM, 2014. pp. 243-247

SILVA, Hadautho Roberto Barros da & ADEODATO, P. J. L. "A Data Mining Approach for Preventing Undergraduate Students Retention". In: THE 2012 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 2012, pp. 1-8.

DIGIAMPIETRI, L. A.; NAKANO, F.; LAURETTO, M. Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso. Revista de Graduação USP, v. 1, n. 1, p. 17-23, 18 jul. 2016.

HOSMER, D.W. and LEMESHOW, S. Applied Logistic Regression. 2nd ed. John Wiley & Sons, Inc. Pp. 156-164, 2000.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321–357, 2002

THAKER, Khushboo; ZHANG, Lei; HE, Daqing; BRUSILOVSKY, Peter. "Recommending Remedial Readings Using Student's Knowledge State". International Conference on Educational Data Mining (EDM), 2020. Fully virtual conference.

YAO, Mengfan; SAHEBI, Shaghayegh; BEHNAGH, Reza Feyzi. "Analyzing Student Procrastination in MOOCs: A Multivariate Hawkes Approach". International Conference on Educational Data Mining (EDM), 2020. Fully virtual conference.