# Qual Técnica de *Learning Analytics* Usar para Prever o Desempenho Acadêmico de Estudantes? Uma Análise Comparativa Experimental com Dados de *MOOCs*

**Welington Veiga, Marcelo de O. C. Machado, Sean W. Siqueira**

Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

Av. Pasteur, 456, Urca – Rio de Janeiro – RJ – Brasil

welington.veiga@gmail.com, marcelo.machado@edu.unirio.br, sean@uniriotec.br

***Abstract.*** *Predicting student academic performance is one of the main research topics in Learning Analytics, for which different techniques have been applied. In order to facilitate the choice of a technique, this study presents a comparative analysis among regression and classification techniques, considering different application scenarios. We used data from MITx/HarvardX containing logs of activities and participation of 15 groups of 12 MOOCs. Results obtained from the performance evaluation metrics suggest the choice of Decision Trees as a technique to build models for regression and a choice between Decision Trees and Support Vector Machines to build models for classification.*

***Resumo.*** *A previsão do desempenho acadêmico de estudantes é um dos principais tópicos de pesquisa em Learning Analytics, para o qual diferentes técnicas foram aplicadas. De modo a facilitar a escolha de uma técnica, este estudo apresenta uma análise comparativa entre técnicas de regressão e classificação, considerando diferentes cenários de aplicação. Foram utilizados dados do MITx/HarvardX contendo logs de atividades e participação de 15 turmas de 12 MOOCs. Os resultados obtidos, a partir das métricas de avaliação de performance, sugerem a escolha de Árvores de Decisão como técnica para criação de modelos para regressão e uma escolha entre Árvores de Decisão e Support Vector Machines para a criação de modelos de classificação.*

## 1. Introdução

O crescente volume de dados gerados em diferentes atividades educacionais, sobretudo em ambientes de educação a distância ou de apoio ao ensino presencial, reforça o interesse de pesquisa em *Learning Analytics* (LA)[1]. De maneira ampla, LA estuda a aplicação de técnicas de análise de dados para compreender e otimizar o processo de ensino-aprendizagem bem como o ambiente em que ele ocorre [Gašević *et al*. 2015]. A previsão de desempenho acadêmico dos estudantes está entre os principais temas de pesquisa em LA [Gašević *et al*. 2016] e pode ser abordada através de perspectivas como a identificação de estudantes sob risco de evasão, tendências de reprovação e até mesmo como a previsão de notas finais.

[1] O termo *Learning Analytics* é utilizado neste trabalho para o contexto geral da área de pesquisa, sem fazer distinção com *Educational Analytics*, *Teaching Analytics* ou *Academic Analytics*.

DOI: 10.5753/cbie.sbie.2019.1391 

Diversas pesquisas relatam a aplicação de técnicas de análise de dados para a previsão de desempenho acadêmico em instituições e cenários específicos. No entanto, essas aplicações frequentemente consideram um escopo reduzido ou um cenário muito específico [Pardo *et al*. 2016; Okubo *et al*. 2017; Almeida *et al*. 2018]. Por exemplo, Yadav e Pal (2012) realizaram uma pesquisa comparativa limitada a testar algoritmos de Árvores de Decisão; Romero *et al.* (2013) compararam técnicas de análise de dados (Regressão Linear, Árvores de Decisão, *Naïve Bayes*) a partir de postagens em fóruns; Jayaprakash *et al.* (2014) compararam a utilização de técnicas (Regressão Logística, Máquina de Vetor de Suporte, Árvores de Decisão e *Naïve Bayes*) para a identificação antecipada de estudantes sob risco de desistência e Naif *et al.* (2017) aplicaram diferentes técnicas (Regressão Logística, Máquina de Vetor de Suporte, Árvores de Decisão e Métodos Bayseanos) a partir de dados socioeconômicos dos estudantes. Nessas pesquisas, a previsão de desempenho acadêmico foi reduzida a um problema de classificação. Além disso, a necessidade de utilização de modelos diferentes de acordo com contextos educacionais distintos também é discutida. Gašević *et al*. (2016) mostram empiricamente, utilizando dados de 4132 alunos de 9 cursos semipresenciais, que não é possível generalizar os resultados obtidos com modelos de predição entre contextos educacionais diferentes. Almeda *et al*. (2018) aplicam técnicas de regressão e classificação analisando dados obtidos tanto em disciplinas oferecidas em *Massive Open Online Course* (MOOC) quanto de creditação obrigatória em cursos de graduação, evidenciando a dificuldade de generalização entre esses contextos.

Na literatura, não foram identificados trabalhos com foco na discussão e análise comparativa entre as diferentes técnicas de LA aplicadas para o problema de previsão de desempenho acadêmico considerando diferentes cursos, cenários e formulações – baseadas tanto em um problema de classificação quanto de regressão. Portanto, objetivando apoiar a escolha adequada de uma técnica, de acordo com o contexto de previsão de desempenho acadêmico de estudantes, este trabalho apresenta uma análise comparativa de técnicas de LA utilizando dados do MITx/HarvardX com *logs* de atividades e participação de 15 turmas de 12 cursos MOOC ofertados entre 2012 e 2013 (versão disponível no período de realização dos experimentos deste trabalho) [Ho *et al*. 2014]. Essa análise foi realizada a partir de uma experimentação que inclui diferentes cenários e formulações. Os resultados obtidos, a partir de métricas de avaliação de performance, sugerem a escolha de técnicas tanto para criação de modelos de regressão quanto de classificação. Assim, novas implementações poderão ser desenvolvidas de acordo com o conhecimento apresentado por este trabalho, vislumbrando, assim, a criação de processos para melhoria de desempenho acadêmico de estudantes.

Além desta introdução, o presente trabalho está organizado da seguinte maneira: a Seção 2 apresenta a fundamentação teórica; a Seção 3 o estudo experimental realizado; a Seção 4 os resultados e a análise comparativa; e a Seção 5 as considerações finais.

## 2. Fundamentação Teórica

O estudo da previsão de desempenho de estudantes se relaciona principalmente com cenários indesejáveis como o abandono e a reprovação. Portanto, pesquisas em LA são estimuladas de forma a criar modelos que permitam a descoberta antecipada e tomada de decisão adequada [Barber e Sharkey 2012].

Os dados utilizados para a previsão de desempenho acadêmico estão frequentemente relacionados a índices de participação e engajamento em AVAs, e consistem em dados coletados automaticamente de acordo com as atividades realizadas, por exemplo, quantidade de vídeos reproduzidos, postagens em fóruns, tempo total no ambiente, entre outros [Romero *et al.* 2013; Jayaprakash *et al.* 2014; Gašević *et al.* 2015; Pardo *et al*. 2016; Almeda *et al*. 2018]. Os dados utilizados em cada estudo dependem do tipo de informação que pode ser extraída, possibilidade de acesso aos dados, extensões e módulos utilizados em cada contexto estudado.

### 2.1. Formulações

A partir dos trabalhos apresentados na literatura, percebe-se que uma forma de analisar os diferentes trabalhos sobre previsão de desempenho acadêmico é quanto a saída esperada. Assim, tal predição pode ser definida a partir de duas formulações, seja como um problema de regressão ou de classificação.

A **Formulação 1 (F1)** agrupa modelos cuja saída esperada é uma previsão direta da nota final que será obtida pelo estudante em uma determinada escala contínua, para os quais são utilizados modelos de regressão [Almeda *et al*. 2018; Pardo *et al*. 2016; Gašević *et al*. 2016]. A saída da F1 é definida em uma escala contínua com notas válidas entre 0 e 1 e atribuindo-se -1 para os estudantes sem nota.

A **Formulação 2 (F2)** agrupa modelos cuja saída esperada é discreta e correspondente a uma categoria dentro de um grupo pré-definido, frequentemente binário, para os quais são utilizados modelos de classificação. A classificação utilizada depende do objetivo do trabalho, por exemplo, provável aprovação ou reprovação, risco de desistência ou de chance de aprovação e até escalas de notas de acordo com rótulos pré-estabelecidos [Romero *et al*. 2013; Jayaprakash *et al.* 2014; Almeda *et al*. 2018]. Neste trabalho, a saída da F2 é definida em duas categorias, a primeira classifica os estudantes **sob risco**, com previsão de reprovação ou desistências e os **concluintes**.

### 2.2. Técnicas de Análise de Dados e Métricas de Avaliação

As técnicas de análise de dados são baseadas em Mineração de Dados e Aprendizagem de Máquina, a partir das quais podem ser construídos modelos parametrizados para problemas específicos. Na literatura de previsão de desempenho acadêmico de estudantes destacam-se os modelos os baseados em **Árvores de Decisão (*Decision Trees* - DT)** [Almeda *et al*. 2018; Pardo *et al.* 2016; Jayaprakash *et al.* 2014; Romero *et al.* 2013], **Aprendizagem Profunda (*Deep Learning* - DL)** [Okubo *et al*. 2017], **Máquinas de Vetor de Suporte (*Support Vector Machines* - SVM)** [Jayaprakash *et al.* 2014], **Métodos Bayseanos (*Bayesian Methods* - NB)** [Jayaprakash *et al.* 2014; Romero *et al.* 2013; Barber e Sharkey 2012], **Regressão Linear (*Linear Regression* - LR)** [Almeda *et al*. 2018; Gašević *et al*. 2016; Romero *et al.* 2013] e **Regressão Logística (*Logistic Regression* - LR)** [Gašević *et al*. 2016; Jayaprakash *et al.* 2014]. Nesse cenário, Shahiri e Husain (2015) apresentam uma revisão sistemática da literatura que corrobora com a indicação das técnicas mais utilizadas citadas anteriormente.

Para avaliar os resultados obtidos, as métricas utilizadas são escolhidas segundo a formulação. Para a F1 predominam: o **erro médio absoluto (MAE)**, que consiste na média da diferença absoluta entre a nota prevista e a real; o **quadrado do erro médio (MSE)**, que consiste no quadrado da diferença entre os resultados previstos e os obtidos;

e o **r² score (R2)** que indica o quanto o resultado previsto pode ser explicado a partir das entradas do modelo. Para a F2 predominam: a **precisão (Prec.)**, que indica a quantidade de resultados positivos corretos dentre o total de positivos previstos para determinada categoria; a **revocação (Rev.)**, que indica o total de positivos encontrados entre o total real de positivos na amostra para determinada categoria; e o **medida-f1 (F1-score)**, que é a média ponderada entre precisão e revocação para determinada categoria.

## 3. Experimentação

A análise comparativa entre os modelos foi realizada por meio de experimentação. Esse método foi escolhido pela possibilidade de execução em laboratório, definindo de maneira precisa e sistemática a aplicação dos métodos e comparação dos resultados [Wholin *et al*. 2012]. A realização de um experimento exige um planejamento cuidadoso por meio de um protocolo que assegure a reprodutibilidade e a interpretabilidade dos resultados obtidos [Munafò *et al*. 2017]. O protocolo definido neste trabalho foi adaptado de Wholin *et al.* (2012), e consiste nas seguintes etapas: (i) planejamento; (ii) análise exploratória dos dados; (iii) implementação dos modelos; (iv) treinamento, execução e coleta de dados; e (v) análise dos resultados. Esta seção discute as quatro primeiras etapas do processo, enquanto a análise comparativa dos resultados é apresentada na próxima seção.

### 3.1. Planejamento

Inicialmente, na etapa de planeamento, foram selecionados os dados utilizados no experimento deste trabalho. *Logs* de atividades e indicadores de participação em AVAs são comumente utilizados na criação de modelos de previsão de desempenho acadêmico [Romero *et al*. 2013; Pardo *et al*. 2016; Almeda *et al*. 2018]. Considerando dados disponíveis de forma aberta e liberados para utilização em pesquisa, foram selecionadas bases de dados disponibilizadas pelo MITx/HarvardX, que representam o *log* consolidado de atividades e participação de 15 turmas de 12 cursos MOOC ofertados entre 2012 e 2013 [Ho *et al*. 2014]. Ainda na etapa de planejamento foram definidas as formulações, as técnicas e métricas utilizadas (Tabela 1).

**Tabela 1. Planejamento do Experimento**

| | Técnicas de Análise de Dados | Métricas | Interpretação |
|---|---|---|---|
| Formulação 1 | Árvores de Decisão | Erro Absoluto Médio (MAE) | Erro absoluto de previsão, em média |
| | Aprendizagem Profunda | Erro ao Quadrado Médio (MSE) | Erro ao quadrado de previsão, em média |
| | Máquinas de Vetor de Suporte Regressão Linear | $R^2$ Score ($R^2$ Score) | Indica quanto o resultado obtido pode ser explicado a partir das características de entrada |
| Formulação 2 | Árvore de Decisão Aprendizagem Profunda | F1 médio (F1-Score) | Média ponderada entre precisão e revogação |
| | Máquinas de Vetor de Suporte Regressão Logística | Precisão (Prec.) | Proporção de positivos reais em relação ao total de positivos |
| | Métodos Bayesianos | Revogação (Rev.) | Proporção entre positivos verdadeiros e falsos negativos |

### 3.2. Análise Exploratória dos Dados

O primeiro passo da análise é compreender o comportamento dos dados disponíveis. Dessa forma, cada variável da base de dados foi investigada para identificação de seu domínio, ausência e inconsistência de valores e correlação com o resultado obtido pelo estudante. No conjunto de dados estudado, existem variáveis obtidas automaticamente

pelo uso do sistema (**Sis.**) e informadas pelos estudantes, dentre as quais algumas foram descaracterizadas (**Desc.**) para evitar a identificação do estudante (Tabela 2).

**Tabela 2. Variáveis Disponíveis para Análise**

| Variável | Sis. | Desc. | Descrição | Valores | Exemplo | IM > 0.1 |
|---|---|---|---|---|---|---|
| Course_id | Sim | Não | Identificação única do curso nas instituições. | Texto | Edx/CB22x/2013 | - |
| User_id | Sim | Sim | Identificação única descaracterizada do estudante. | Texto | MHxPC130442623 | - |
| Registered | Sim | Não | Indica se o usuário está registrado no curso. Sim para todos os registros. | Sim/Não | 1 | - |
| Viewed | Sim | Não | Indica se o estudante visualizou o curso durante sua realização. | Sim/Não | 1 | Sim |
| Explored | Sim | Não | Indica se o estudante visualizou pelo menos metade dos capítulos do curso. | Sim/Não | 0 | Sim |
| Certified | Sim | Não | Indica se o estudante obteve o certificado no fim do curso. | Sim/Não | - | - |
| Final_cc_name_DI | - | Sim | País/Região do estudante. Parte computada automaticamente pelo IP, parte informada pelo estudante. | Texto | South America | Não |
| LoE | Não | Não | Grau de escolaridade do estudante. | Texto | Master | Sim |
| YoB | Não | Não | Ano do nascimento. | Inteiro | 1989 | Sim |
| Gender | Não | Não | Gênero do estudante. | F/M/O | F | Não |
| Grade | Sim | Não | Nota final do estudante ausente ou entre 0 e 1. | Decimal | 0.87 | - |
| Start_time_DI | Não | Sim | Registro do estudante no curso. | Data | 12/18/12 | Não |
| Last_event_DI | Sim | Sim | Última interação do estudante no curso. | Data | 11/17/13 | Sim |
| Ndays_act | Sim | Não | Número de dias que o estudante interagiu com o curso. | Inteiro | Inteiro | Sim |
| Nplay_video | Sim | Não | Número de reproduções de vídeos. | Inteiro | 21 | Sim |
| Nchapters | Sim | Não | Números de capítulos do curso acessado. | Inteiro | 10 | Sim |
| Nforum_posts | Sim | Não | Número de postagem em fóruns. | Inteiro | 12 | Não |
| Roles | Sim | Não | Papel no ambiente, identifica os tutores e professores removidos da base. | Texto | - | - |
| Inconsistente_flag | Sim | Não | Indica existência de alguma inconsistência no registro. | Sim/Não | 1 | - |

Analisando os dados em relação aos resultados dos alunos por curso, é possível verificar que há um número maior de estudantes sem notas ou com notas iguais a zero do que com notas superiores a zero (Figura 1). Esse resultado reforça estudos que indicam que em MOOC o objetivo dos estudantes não é necessariamente ser aprovado ou obter um certificado, o que pode explicar o índice de desistência superior ao de disciplinas tradicionais [Liyanagunawardena *et al*. 2017; Gašević *et al*. 2016].

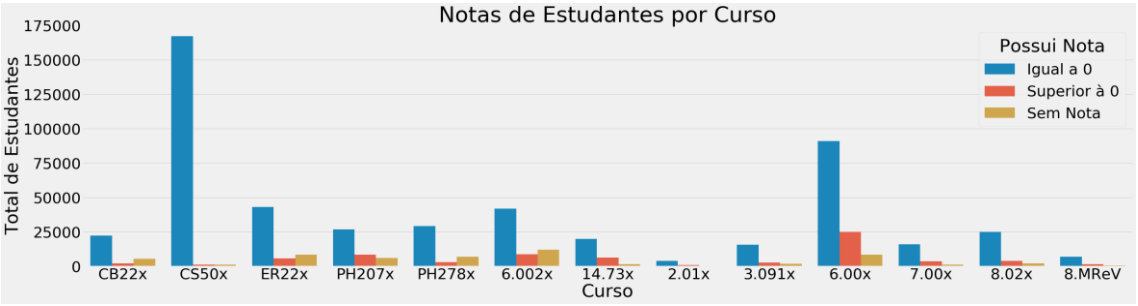


**Figura 1. Quantidade de estudantes por faixas de nota final por curso**

Para a análise das variáveis em relação às notas, é necessário considerar apenas as notas superiores a zero para que seja possível observar outros padrões além da predominância desses valores. A partir das variáveis qualitativas é possível reforçar a relação entre a exploração do conteúdo com o desempenho acadêmico dos estudantes, por exemplo, pela comparação da variação da nota final de toda a população

(matriculados), dos que visualizaram o curso (visualizou) e dos que exploraram o conteúdo do curso no mínimo até a metade (exploração > 50%) (Figura 2).

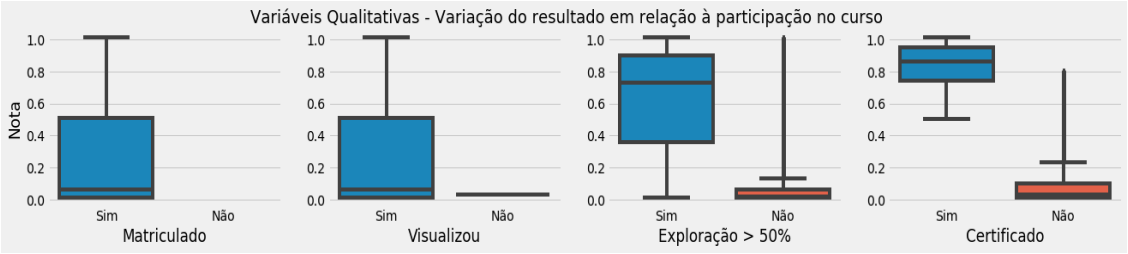


**Figura 2. Variação dos resultados em função de variáveis qualitativas**

Em relação às variáveis quantitativas sobre a atividade dos estudantes no ambiente, se relacionadas com a nota final, é possível verificar que há uma correlação entre o resultado obtido e quantidade de dias de interação, vídeos exibidos, capítulos explorados e postagens em fóruns no ambiente (Figura 3).

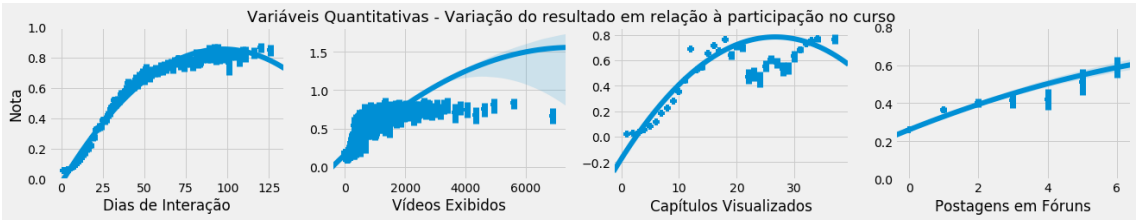


**Figura 3. Variação dos resultados em função de variáveis quantitativas**

A partir da análise das variáveis relacionadas às atividades dos alunos é possível verificar padrões em relação aos resultados dos estudantes, o que intuitivamente indica que as técnicas de análise de dados podem obter bom desempenho na construção de modelos que expressem esses padrões.

### 3.3. Implementação

Na literatura, de acordo com a formulação e os objetivos de cada pesquisa, as técnicas de análise de dados aplicadas são implementadas, parametrizadas e adaptadas. Assim, foram utilizadas as bibliotecas Scikit-learn e Keras para a linguagem de programação Python 3.5, que oferecem implementações e uma parametrização padrão conveniente para cada modelo. As variáveis categóricas e numéricas foram normalizas para utilização nos modelos. Todo o código utilizado está disponível publicamente[2].

### 3.4. Treinamento, Execução e Coleta dos Dados

Para oferecer uma visão abrangente do comportamento de cada modelo foram definidos 3 cenários simulando variações em relação à presença de dados inconsistentes e à realização seleção de variáveis específicas com maior relevância para solução do problema. Dessa forma, os cursos foram divididos aleatoriamente em 3 cenários com 5 cursos cada: **Cenário 1**, onde todas as variáveis e todos os estudantes dos cursos foram utilizados; o **Cenário 2**, apenas com os estudantes com dados consistentes; e o **Cenário 3** com todos os estudantes, porém apenas as variáveis selecionadas de acordo com sua relação com a saída desejada. Para a seleção de variáveis no Cenário 3 foi utilizada a

[2] https://github.com/marcelomachado/la-perfomance-mooc

métrica **informação mútua (IM)** que indica a relação entre uma variável e a saída esperada, mesmo que em padrões não lineares [Vergara e Estévez 2014].

Após o pré-processamento dos dados, os modelos foram aplicados para cada formulação em cada um dos cursos separadamente, onde parte dos dados disponíveis foram separados a priori para avaliação. As métricas de interesse foram coletadas de forma automática pelo *script* responsável pela execução dos testes.

## 4. Resultados e Análise Comparativa

A partir do experimento realizado, os dados coletados foram consolidados de forma estruturada por cenário, modelo e métrica para a F1 (Tabela 3) e para a F2 (Tabela 4).

**Tabela 3 - Resultados e métricas coletadas no experimento para a Formulação 1**

| Formuláção 1 - Regressão: Previsão do Valor da Nota. | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo | | Árvores de Decisão | | | Aprendizagem Profunda | | | Máquinas Vetor de Suporte | | | Regressão Linear | | |
| Tempo (Treino/Teste) | | 247.5s / 0.4s | | | 130.8s / 1.7s | | | 143.1 / 0.5 | | | 248s / 0.4s | | |
| Métricas | | MAE | MSE | $R^2$score | MAE | MSE | $R^2$score | MAE | MSE | $R^2$score | MAE | MSE | $R^2$score |
| Cenário 1 | n=270757 | **0,087** | 0,079 | 0,193 | 0,132 | 0,098 | 0,117 | 0,148 | 0,070 | 0,197 | 0,141 | 0,068 | 0,270 |
| Cenário 2 | n=175529 | **0,021** | 0,007 | 0,779 | 0,033 | 0,008 | 0,724 | 0,067 | 0,009 | 0,724 | 0,036 | 0,007 | 0,773 |
| Cenário 3 | n=148408 | **0,136** | 0,125 | 0,168 | 0,201 | 0,107 | 0,260 | 0,191 | 0,108 | 0,269 | 0,256 | 0,106 | 0,240 |

**Tabela 4 - Resultados e métricas coletadas no experimento para a Formulação 2**

| Formulação 2 - Classificação: Previsão de Estudantes Sob Risco | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo | | Árvores de Decisão | | | Aprendizagem Profunda | | | Máquinas Vetor de Suporte | | | Regressão Logística | | | Naïve Bayes | | |
| Tempo (Treino/Teste) | | 33.7s / 2.5s | | | 115.2s / 10s | | | 62.5s / 2.5s | | | 88.1s / 2.2s | | | 34.8s / 2.1s | | |
| Métricas | | F1 | Prec. | Rev. | F1 | Prec. | Rev. | F1 | Prec. | Rev. | F1 | Prec. | Rev. | F1 | Prec. | Rev. |
| Cenário 1 | n=270757 | 0,737 | 0,782 | 0,697 | 0,029 | 0,015 | 0,588 | **0,764** | 0,761 | 0,770 | 0,745 | 0,761 | 0,735 | 0,452 | 0,307 | 0,982 |
| Cenário 2 | n=175529 | 0,818 | 0,850 | 0,791 | 0,043 | 0,023 | 0,596 | **0,821** | 0,849 | 0,797 | 0,815 | 0,839 | 0,795 | 0,544 | 0,411 | 0,975 |
| Cenário 3 | n=148408 | 0,841 | 0,843 | 0,840 | 0,085 | 0,044 | 0,998 | **0,862** | 0,829 | 0,900 | 0,849 | 0,829 | 0,872 | 0,713 | 0,564 | 0,995 |

A partir da consolidação dos dados coletados durante o experimento, é possível comparar o desempenho de cada modelo sob diferentes aspectos. Primeiro, quanto à formulação, fica evidente como a mudança na definição da saída impacta na qualidade dos resultados, pois as mesmas técnicas se comportaram de forma diferente quando a saída foi tratada como regressão (F1) ou como classificação (F2).

Quanto aos 3 cenários avaliados, no entanto, a análise pode ser dividida em duas perspectivas. Primeiro, é possível verificar que a remoção de dados inconsistentes interfere nos resultados, com uma melhora média de 69% no MAE para a F1 e 28% no F1-Score para a F2. Além disso, a seleção de variáveis específicas não implica necessariamente em melhoria dos resultados, uma vez que para a F1 o MAE piorou 54% em média, e o para a F2 o F1-Score melhorou 11%, indicando que a escolha das variáveis impacta os modelos de forma diferente considerando cada formulação. Por fim, a divisão em cenários não apresentou impacto considerável na comparação de performance das técnicas, de modo que as técnicas com melhores resultados foram consistentes, sendo melhores entre os cursos e cenários.

Quanto às técnicas, é importante observar a performance dos resultados baseada no MAE (F1) e no F1-Score (F2). Utilizando uma escala que vai de 0 (pior

resultado) a 1 (melhor resultado) é possível verificar que: na F1, as DT obtiveram resultados superiores às demais técnicas com maior estabilidade entre a qualidade dos resultados para cada curso entre os cenários; e, na F2, os resultados obtidos com DT, SVM e LR ficaram próximos tanto em relação à qualidade quanto à estabilidade (Figura 4).

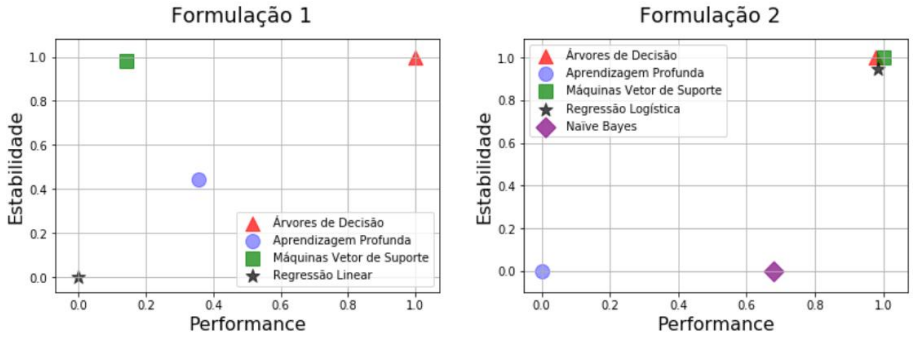


**Figura 4. Comparação entre a performance dos modelos.**

Um outro fator de decisão é o tempo necessário para treinamento dos modelos e para realização de previsões. A Figura 5 apresenta uma visão comparativa dos modelos de acordo com o tempo total necessário para treinamento e para predição durante o experimento em uma escala de 0 (pior tempo) a 1 (melhor tempo). Para a F1 as técnicas que obtiveram melhor tempo para previsão foram as DT e LR, enquanto a DL foi o modelo com menor tempo de treinamento e as SVM tiveram bons tempos de previsão e treinamento. Para a F2 as técnicas de DT e NB oferecem os melhores tempos tanto para treinamento quanto para predição, enquanto as SVM possuem um resultado bom para predição, porém moderado para treinamento.

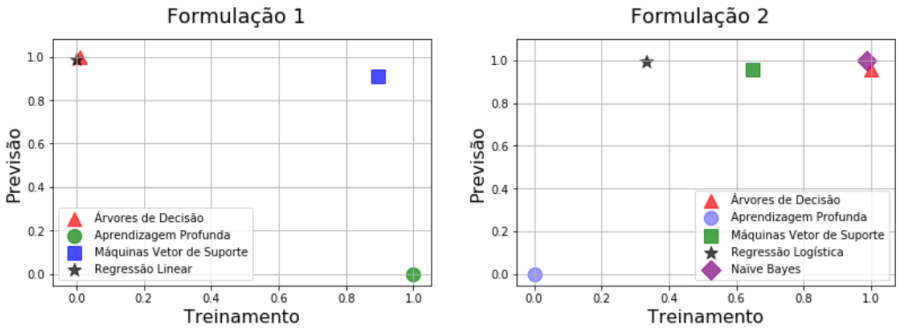


**Figura 5. Comparação tempo gasto para treinamento e previsão.**

Os resultados obtidos a partir das métricas de avaliação de performance sugerem a escolha de **DT como técnica para criação de modelos para regressão e uma escolha entre DT e SVM para a criação de modelos de classificação**. Além disso, os resultados reforçam que outros fatores devem ser considerados na escolha de uma técnica de análise de dados em detrimento de outra, como a estabilidade entre diferentes cenários, importância da formulação, limpeza dos dados inconsistentes, seleção adequada de variáveis, e os tempos necessários para treinamento e previsão.

## 5. Considerações Finais

A previsão de desempenho acadêmico de estudantes é um importante problema de pesquisa para o qual a escolha da técnica de análise de dados é uma tarefa complexa.

Nesse contexto, o presente trabalho apresentou uma análise comparativa experimental de técnicas para classificação e regressão, identificadas na literatura. O experimento realizado utiliza dados disponibilizados pelo MITx/HarvardX, através de *logs* de atividades em AVAs, consolidando a participação de 15 turmas de 12 cursos MOOC ofertados entre 2012 e 2013. Esses dados representam o tipo de aplicação predominante para a previsão de desempenho acadêmico. Embora os estudos como esses dados possam não capturar as reais causas de um baixo desempenho ou motivação da desistência de um estudante, padrões de comportamento dos estudantes nesses ambientes podem ser importantes fontes de informação para identificar esses estudantes antecipadamente [Barber e Sharkey 2012]. Também é importante entender que as previsões realizadas possuem significado se alinhadas com o processo de ensino-aprendizagem e teoria pedagógica aplicada [Jayaprakash *et al*. 2014; Gašević *et al*. 2016], reforçando a importância do entendimento dos modelos aplicados.

Diferentes trabalhos apontam que modelos de LA desenvolvidos para um curso específico não podem ser estendidos para outros contextos [Almeda *et al*. 2018; Gašević *et al*. 2016], no entanto essa limitação não foi observada nos resultados obtidos, o que pode indicar a existência de contextos mais amplos do que os sugeridos (como uma plataforma MOOC), em que os mesmos modelos possam ser replicados.

Os resultados apresentados oferecem um importante ponto de partida para compreender as características das técnicas para o problema de previsão de desempenho acadêmico, apoiando uma escolha adequada segundo o contexto. Como limitação da pesquisa, pode-se destacar o fato dos dados utilizados serem provenientes de uma mesma fonte de dados, MITx/HarvardX, o que pode implicar em abordagens didático-pedagógicas semelhantes, bem como padrões de comportamentos de usuários similares. Como trabalhos futuros, novos experimentos e estudos de caso com outras bases de dados e contextos educacionais podem ampliar a compreensão tanto do problema de previsão de desempenho acadêmico quanto das técnicas aplicáveis a cada contexto. Ainda, é pretendida a criação de instanciações de soluções de acordo com as técnicas selecionadas por este trabalho, com objetivo de utilização em AVAs, para apoiar a tomada de decisão preventiva e vislumbrando apoiar a melhoria de desempenho acadêmico dos estudantes.

## Referências

Almeda, M. V., Zuech, J., Baker, R. S., Utz, C., Higgins, G., Reynolds, R. (2018). Comparing the Factors that Predict Completion and Grades among For-Credit and Open/MOOC students in Online Learning. *Online Learning*, 22(1), 1-18.

Barber, R., Sharkey, M. (2012). Course correction: Using Analytics to Predict Course Success. *In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, (pp. 259-262). ACM.

Gašević, D., Dawson, S., Rogers, T., Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64-71.

Ho, A., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., Chuang, I. (2013) HarvardX and MITx: The First Year of Open Online Courses, fall 2012-summer 2013. *HarvardX and MITx Working Paper No. 1*. Available at SSRN: https://ssrn.com/abstract=2381263 .

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics, 1*(1), 6–47.

Liyanagunawardena, T. R., Parslow, P., Williams, S. A. (2017). Exploring 'success' in MOOCs: Participants' perspective. *In Massive Open Online Courses and Higher Education*, (pp. 106-122). Routledge.

Munafò, R.; Nosek, B.; Bishop, D.; Button, K.; Chambers, C.; Percie du Sert, N., Simonsohn, U. & Wagenmakers, E. e J. Ware, Jennifer & P. A. Ioannidis, John. (2017). A manifesto for reproducible science. Nature Human Behaviour. 1. 0021.

Naif, A. D., Rabeeh R. A., Miltiadis, A. A., Farhat D. L, Abbas, J. S. A. (2017). Predicting student performance using Advanced Learning Analytics. *In Proceedings of the 26th International Conference on World Wide Web Companion,* (pp. 415-421).

Okubo, F., Yamashita, T., Shimada, A., Ogata, H. (2017) A Neural Network Approach for Students' Performance Prediction. *In Proceedings of the 7th International Conference on Learning Analytics & Knowledge,* (pp. 598-599). ACM.

Pardo, A., Mirriahi, N., Martinez, R., Jovanovic, J., Dawson, S., & Gašević, D. (2016). Generating actionable predictive models of academic performance. *In Proceedings of the 6th International Conference on Learning Analytics & Knowledge,* (pp. 474-478). ACM.

Romero, C., López, M. I., Luna, J. M., Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68, 458-472.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1), 175-186.

Yadav, S.K., Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal*. (ISSN: 2221-0741), Vol. 2, No. 2, 51-56, 2012.