# Análise de Regressão Aplicada a Previsão de Reprovação de Alunos em Plataforma de Ensino a Distância

*Analysis of Regression Applied to the Reproducibility Forecast of Students in Distance Learning Platform*

**Francisco de Assis de Araújo** [1] orcid.org/0000-0002-6052-5330
**Rodrigo Lins Rodrigues** [2] orcid.org/0000-0002-4521-3806

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.
[2] Departamento de Educação, Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brasil.

**E-mail do autor principal:** faa@ecomp.poli.br

**Resumo**

Um dos principais problemas enfrentados no Ensino a Distância são os riscos de reprovação e evasão de alunos. Com o objetivo de auxiliar Professores e gestores nessa modalidade de ensino, este trabalho demonstra resultados das aplicações práticas de técnicas estatísticas e mineração de dados para previsão de reprovação de Alunos através da Análise de Regressão Logística que demonstrou sua eficácia através de excelentes índices de desempenho em três modelos de dados utilizados, índices estes que foram considerados estatisticamente iguais através da Análise de Variância (ANOVA) aplicada ao comparar os índices de desempenho dos modelos de Regressão gerados. Através dos índices de significância das variáveis selecionadas em cada modelo é possível identificar os meios de interação que mais contribuem com o desempenho do aluno, auxiliando no combate a reprovação.

**Palavras-Chave:** Previsão de Reprovação; Análise de Regressão; EAD;

***Abstract***

*One of the main problems faced in Distance Learning is the risks of student disapproval and avoidance. With the objective of assisting teachers and managers in this teaching modality, this paper demonstrates the results of the practical applications of statistical techniques and data mining to predict student disapproval through Logistic Regression Analysis that demonstrated its effectiveness through excellent performance indices in three data models used, which were considered statistically equal by the Analysis of Variance (ANOVA) applied when comparing the performance indices of the Regression models generated. Through the indices of significance of the variables selected in each model, it is possible to identify the means of interaction that contribute most to the student's performance, helping to combat failure.*

***Key-words:*** *Forecast of Reprobation; Regression Analysis; EAD;*

 http://dx.doi.org/10.25286/repa.v3i3.954

## 1 Introdução

A Educação a Distância (EAD) no Brasil tem se consolidado com diversos estudantes optando por essa modalidade de ensino para ampliar suas formações e realização profissional. Um dos principais diferenciais desta modalidade de ensino é a grande quantidade de dados gerada pelas interações nas plataformas de suporte online, conhecidas como AVA ou Ambientes Virtuais de Aprendizagem que abre novas possibilidades para pesquisas buscando compreender os processos de aprendizagem por meio das interações de alunos e professores. De acordo com Cechinel et al, algumas áreas de pesquisas surgiram nos últimos anos com intuito de auxiliar em questões como essas **[1]**.

De acordo com Cristobal et al, a Mineração de Dados Educacionais (do inglês Educacional Data Mining - EDM) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional **[2]**.

Baker et al., aponta a possibilidade de identificação de estudantes com alto risco de evasão e reprovação a partir de modelos automáticos como um dos potenciais problemas a serem atacados pela comunidade brasileira que atua na área de mineração de dados educacionais **[7]**.

Conforme CECHINEL et al, a eficácia e a eficiência de estudantes têm frequentemente sido associadas a diferentes medidas de suas interações dentro dos Ambientes Virtuais de Aprendizagem-AVA, medidas estas que normalmente possuem uma alta correlação com o sucesso do aprendizado dos alunos **[1]**.

As interações dos alunos e professores com os ambientes virtuais de aprendizagem (AVA) proveem os dados que alimentam as pesquisas nessas áreas e possibilitam a descoberta de novos conhecimentos.

Este trabalho tem como objetivo estimar um modelo de relação existente entre a reprovação do aluno e as interações quantificadas através dos logs de interações com o ambiente de ensino, tendo como resultado um Modelo Regressão Logística capaz de prever o risco de reprovação do aluno.

Este trabalho está organizado da seguinte maneira: A Seção 2 apresenta alguns trabalhos relacionados que demonstram semelhanças com o trabalho atual, a Seção 3 apresenta um breve histórico sobre Análise de Regressão e o Modelo de Regressão Logístico, a Seção 4 apresenta o objetivo do trabalho e os modelos de dados utilizados no experimento de Análise de Regressão, a Seção 5 a base de dados através de algumas análises descritivas das variáveis, a Seção 6 o pré-processamento e limpeza dos dados, a Seção 7 apresenta o desenvolvimento da modelagem através da Análise de Regressão logística, na Seção 8 os Modelos Logísticos gerados são avaliados através de seus índices de desempenho e na Seção 9 são apresentadas as conclusões.

## 2 Trabalhos Relacionados

Em diversos trabalhos relacionados à EAD, observa-se que o sucessos dos alunos em ambiente de AVA está diretamente relacionado às diferentes medidas de interações com o ambiente. Por exemplo, Murray et al., observou que estudantes que apresentaram as mais altas taxas de acesso aos conteúdos no AVA obtiveram desempenhos mais satisfatórios nas avaliações **[4]**.

Dickson et al**.**, descobriu que o número total de cliques dados por estudantes é fortemente correlacionado com as suas notas finais em um curso **[5]**. Manhães et al**.**, utilizou técnicas de mineração de dados para prever a evasão de estudantes em cursos presenciais da Escola Politécnica da Universidade do Rio de Janeiro **[6]**.

Cechinel et al, descreve resultados da aplicação de técnicas de aprendizado de máquina para demonstrar a viabilidade de utilizar apenas a quantidade de interações dos alunos para gerar predições razoavelmente precisas e que a introdução de atributos derivados das contagens (e.g. médias) é útil para previsões mais precisas quando a quantidade de dados é esparsa **[1]**.

Esses estudos assemelham-se ao presente trabalho através da utilização de atributos de contagens de interações para treinamento do modelo de predição, contudo diferente de alguns trabalhos, este busca desenvolver uma modelagem capaz de prevê a reprovação do aluno no intuito de colaborar com os Professores no combate a essa reprovação e consequentemente a evasão no ambiente de ensino.

## 3 Análise de Regressão

Segundo Batista, a expressão "regressão" em estatística significa a dependência funcional entre duas ou mais variáveis aleatórias, correspondendo em termos matemáticos à obtenção de uma função que melhor represente a dependência entre aquelas variáveis. O Modelo *logit* foi inserido na terminologia estatística médica por Berkson, em 1944, que batizou o referido termo, por analogia com o modelo desenvolvido por Bliss em 1934, designado por *probit*. O Modelo de Regressão Logístico, também denominado por Modelo *logit,* é especialmente adaptável aos casos em que existe uma variável dependente binária ou dicotômica **[8].**

Através da Regressão Logística, avaliamos o relacionamento entre uma variável dependente e diversas outras variáveis intituladas de independentes, sendo a variável dependente ou resposta $Y_j$ binária, essa variável binária pode assumir os valões $Y_j=0$ ou $Y_j=1$ que em nosso caso significa "aprovado" e "reprovado", respectivamente. Neste caso, "sucesso" é o evento de interesse e significa prever se o aluno será reprovado conforme as suas interações com o ambiente educacional.

De acordo com Baker et al, Uma forma de realizar essa previsão é aplicar o modelo de Regressão Logística, muito popular em EDM para predições binárias **[7]**. Segundo Batista **[8]**, o Modelo Logístico não tem muitas exigências de pressupostos e é usado para prever a probabilidade de eventos binários ocorrerem em que: se a probabilidade for > 0,5 a previsão é de que o resultado do evento seja 1 (em nosso caso significa reprovado) e <= 0,5 a previsão é de que o resultado do evento seja 0 (em nosso caso significa aprovado). Um modelo de regressão Logística segue a equação **(1)**.

$$\boldsymbol{logit}(\boldsymbol{p_j}) = \boldsymbol{ln}\left(\frac{\boldsymbol{p_j}}{\boldsymbol{1-p_j}}\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{x}_1 + \cdots + \boldsymbol{\beta}_n\boldsymbol{x}_n \quad \textbf{(1)}$$

Em que $\mathbf{p}_j$ indica a probabilidade de ocorrência, $x_1 \ldots x_n$ representa o vetor de variáveis explicativas (ou independentes) e $\beta_o$ e $\beta_x$ indicam os coeficientes do modelo **[8]**.

## 4 Experimento Realizado

O experimento deste trabalho se dera através de uma base de dados referente às interações de Alunos em uma plataforma de Ensino a Distância (EAD) com os diversos artefatos educacionais disponibilizados durante os cursos de Pedagogia, Administração e Biologia através do primeiro ao oitavo períodos letivos, com o objetivo de descobrir um modelo de dados ideal para previsão de reprovação de alunos através de Regressão Logística. Neste experimento foram utilizados os modelos de dados Base Geral, Base recortada por Cursos e Base recortada por períodos, as variáveis participantes foram selecionadas através do método de Regressão *Stepwise*. Para testar se existe diferença entre os desempenhos dos três modelos gerados, os índices de desempenho dos modelos de predição foram comparados entre si através da Análise de Variância (*ANOVA*), assim como a comparação visual através da Curva ROC de cada modelo gerado.

## 5 Análise Descritiva dos Dados

A Base de dados disponibilizada no formato original xls, refere-se às interações de 1738 Alunos em uma plataforma de Ensino a Distância (EAD) com os diversos artefatos educacionais disponibilizados durante os cursos de Administração, Biologia, Letras e Pedagogia. Esta base contém 30217 linhas de dados e 39 variáveis, sendo 6 variáveis ordinais, 32 variáveis discretas, 1 variável continua.

Também fazem parte da base de dados as variáveis ordinais Curso, Período, Semestre, Id do Aluno, Id da Disciplina, Nome da disciplina e a variável contínua DESEMPENHO além das variáveis candidatas a preditoras descritas a seguir no quadro 1.


http://dx.doi.org/10.25286/repa.v3i3.954

**Quadro 1 -** Descrição das variáveis

| Variável | Descrição sobre as variáveis |
|---|---|
| **VAR01** | Quantidade de diferentes locais (IP´s) a partir dos quais a(o) aluna(o) acessou o ambiente. |
| **VAR02** | Quantidade de mensagens enviadas por aluna(o) às(os) Professoras(es) pelo ambiente. |
| **VAR03** | Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente. |
| **VAR04** | Quantidade geral de mensagens enviadas pela(o) aluna(o) dentro do ambiente. |
| **VAR05** | Quantidade geral de mensagens recebidas pela(o) aluna(o) dentro do ambiente. |
| **VAR06** | Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo "tira-dúvidas". |
| **VAR07** | Quantidade de postagens no "Fórum tira dúvidas"; |
| **VAR08** | Quant. de postagens de um(a) aluno(a) em fóruns que foram respondidas por outros(as) alunos(as). |
| **VAR09** | Quantidade de postagens de um(a) aluno(a) em fóruns que foram respondidas pelo(a) professor(a) ou tutor(a). |
| **VAR10** | Quantidade de colegas diferentes para quem o(a) aluno(a) enviou mensagens dentro do ambiente. |
| **VAR12** | Quantidade de visualizações da aba "Conteúdo" do curso, onde constam os arquivos com o conteúdo programático do curso |
| **VAR13** | Horário que mais realizou atividades; |
| **VAR14** | Turno do dia em que realizou mais atividades. |
| **VAR16** | Quantidade de atividades entregues por um(a) aluno(a) fora do prazo, por disciplina; |
| **VAR17** | Tempo médio entre a abertura da atividade e sua submissão; |
| **VAR18** | Quantidade de leituras feitas ao fórum (*pageviews*); |
| **VAR20** | Quantidade de respostas ao tópico principal (refazer opinião em fórum); |
| **VAR21** | Quantidade de pageviews ao quadro de notas; |
| **VAR22** | Quantidade de vezes que o aluno visualiza o (*Cheklist* Atividades) |
| **VAR23** | Quantidade de visualizações de notas por atividade; |
| **VAR24** | Média semanal da quantidade de acessos de um(a) aluno(a) ao ambiente. |
| **VAR25** | Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno; |
| **VAR31** | Quantidade de acessos do(a) aluno(a) ao ambiente. |
| **VAR31b** | Quantidade de dias distintos que o aluno entrou na disciplina |
| **VAR31c** | Quantidade de dias distintos que o aluno entrou na plataforma |
| **VAR32a** | Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Manhã). |
| **VAR32b** | Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Tarde). |
| **VAR32c** | Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Noite). |
| **VAR32d** | Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Madrugada). |
| **VAR33** | Quantidade de atividades entregues por um(a) aluno(a) no prazo, por disciplina; |
| **VAR34** | Quantidade geral de postagens de um(a) aluno(a) em fóruns. |
| **VAR35** | Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns. |

Para um melhor entendimento, fez-se necessário a realização de algumas análises descritivas, para isto o gráfico de barras apresentado na figura 10 mostra as médias de acessos por aluno em cada meio através do qual o aluno interagiu com o ambiente de ensino, portanto a média do total de acessos por aluno (Var31) foi de 1470, o meio através do qual o Aluno menos interagiu com o ambiente foi em "Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo tira-dúvidas" (VAR06), "Quantidade de postagens no Fórum tira dúvidas" (VAR07), "Quantidade de respostas ao tópico principal (refazer opinião em fórum)" (VAR20) e "Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns" (VAR35) o meio através do qual os alunos mais interagiram com o ambiente foi em "quantidade de mensagens recebidas dentro do ambiente" (Var05) com uma média de 667 mensagens por aluno, porém a média de mensagens enviadas pelos alunos (Var03) foi de 284 mensagens, o turno em que o aluno mais acessou o ambiente foi o turno da noite (Var32c) com uma média de 544 acessos e o turno que menos o aluno acessou foi o turno da madrugada (Var32d) com uma média de 32 acessos, o meio onde houve menos interações foi em "Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo tira-dúvidas" (Var06) com uma média de 0,48 por aluno, foi realizado uma análise da média de acessos para o DESEMPENHO e constatado que os DESEMPENHOS >4 tiveram uma média de 1315 acessos e os DESEMPENHOS <=4 tiveram uma média de 489 acessos. Através do gráfico de barras apresentado na figura 1, visualiza-se um comparativo ente as médias de acessos dos alunos em cada ambiente de interação através das variáveis candidatas a preditoras.

## 6 Pré-Processamento dos Dados

Esta etapa tem como objetivo melhorar a qualidade dos dados e a eficiência do processo de

mineração através da remoção de dados ruidosos, valores faltantes e dados inconsistentes oriundos da coleta dos mesmos. Na base de dados original no formato .xls, foi efetuado a mudança do padrão numérico Brasileiro para o padrão americano, substituindo a virgula (separador decimal) por ponto. A variável "Quantidades de Time Out" (VAR28), mesmo constando no dicionário de dados, foi excluída das análises devido inexistência na base de dados original. Conforme quadro 2, não sendo constatado interações no Semestre 2009.2, foi considerado como dados faltantes ou inexistentes e o referido semestre foi excluído da base de dados.

Conforme quadro 3, referente ao sumário da variável DESEMPENHO, constatando-se o limite superior igual a 11, considerado como inconsistência o referido DESEMPENHO foi reduzido para 10.

Através da variável DESEMPENHO, foi criada a variável DESEMPENHO_BINÁRIO, através do seguinte critério: A variável criada recebeu "0" para aprovado quando DESEMPENHO >= 4 e "1" para reprovados quando DESEMPENHO < 4.

**Quadro 2 -** Totais de Alunos e interações por Semestre

| Semestre | Alunos | Interações |
|---|---|---|
| **2009.2** | 50 | 0 |
| **2010.1** | 170 | 54349 |
| **2010.2** | 259 | 211757 |
| **2011.1** | 276 | 183265 |
| **2011.2** | 260 | 172672 |
| **2012.1** | 195 | 100904 |
| **2012.2** | 24 | 3069 |
| **2013.2** | 543 | 583463 |
| **2014.1** | 530 | 241758 |
| **2014.2** | 1295 | 524079 |
| **2016.1** | 1294 | 480342 |

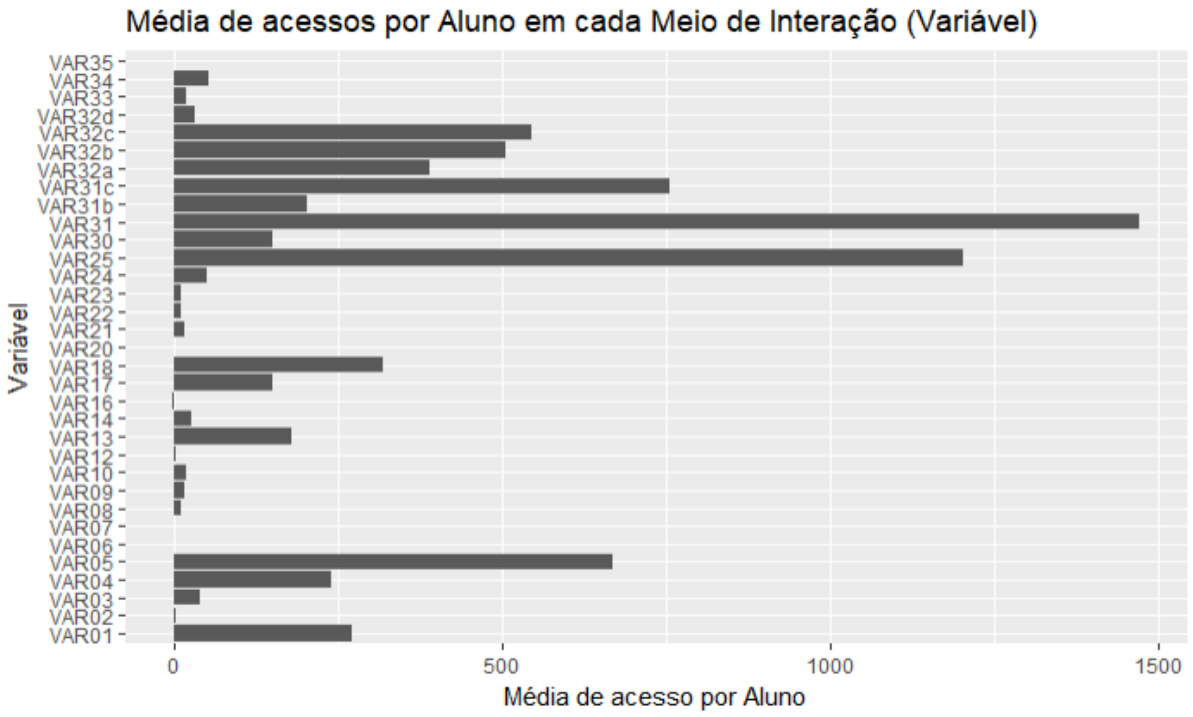



**Figura 1:** Média de acessos por aluno ao meio de interação


http://dx.doi.org/10.25286/repa.v3i3.954

**Quadro 3 -** Sumarização da variável DESEMPENHO

| | |
|---|---|
| Min. | 0.000 |
| 1st Qu. | 1.000 |
| Median | 5.000 |
| Mean | 4.298 |
| 3rd Qu. | 7.000 |
| Max. | 11.000 |

## 7 Desenvolvimento da Modelagem

Neste trabalho, estamos interessados em encontrar o melhor modelo de dados para o desenvolvimento de um modelo de Regressão Logístico capaz de realizar previsão de reprovação dos Estudantes com excelentes índices de precisão, para isto buscou-se descobrir o modelo de dados com melhor potencial de eficácia para o Modelo de Regressão. Os modelos de dados utilizados foram a Base Geral sem recortes e os recortes realizados através das variáveis Curso e Período. O objetivo deste recorte foi identificar as melhores configurações para a construção do modelo: (1) Modelo Genérico com todos os alunos, (2) Modelo por Período e (3) Modelo por Curso.

Em cada um dos 13 modelos de dados analisados, 1 modelo através da Base Geral sem recortes, 4 através dos recortes por Curso e 8 através dos recortes por Período, foi aplicado o método de Regressão *stepwise* para adicionar sistematicamente a variável mais significativa ou remove a variável menos significativa e determinar um melhor subconjunto de variáveis preditoras para o modelo de dados. O quadro 5 apresenta as variáveis selecionadas a partir do modelo de dados Base Geral sem recortes, e seus índices de significância no modelo.

Conforme os modelos de dados, as análises realizadas geraram 13 Modelos de Regressão Logística, a partir desses modelos foram geradas as matrizes de confusão e extraídos os índices de desempenho do modelo Base Geral conforme quadro 6, as médias dos índices de precisão dos modelos por Período conforme quadro 7 e as médias dos índices precisão dos modelos por Curso conforme quadro 8. Conforme Kuhn M (2013), o quadro 4 descreve os índices de desempenho usados na avaliação dos modelos. A seguir serão apresentados três cenários nos quais foram analisados os três modelos de dados através da Análise de Regressão Logística e definidos os modelos de previsão a serem comparados entre si através de seus índices de desempenho, assim como a Curva ROC que, segundo Batista, baseada na taxa de verdadeiros positivos TPR e na taxa de falsos positivos FPR, descreve graficamente o desempenho do sistema classificador binário **[8].**

**Quadro 4 -** Índices de desempenho usados na avaliação dos modelos

| | |
|---|---|
| ***Accuracy*** | Descreve com que frequência o classificador está correto |
| ***Kappa*** | Diferença entre a precisão e a taxa de erro nulo |
| ***Sensitivity*** | Razão entre *True* Positivos para o verdadeiro sim geral |
| ***Specificity*** | Razão de positivos verdadeiros para o total real não |
| ***Precision*** | Razão de positivos verdadeiros para o previsto global sim. |
| ***Prevalence*** | Relação entre o sim real e o número total de instâncias. |
| ***Balanced_Accuracy*** | Média de sensibilidade e especificidade |

### 7.1 Cenário "Modelo Geral"

Neste modelo de dados foi utilizado a base sem recortes, as variáveis Selecionadas e seus índices de significância estão descritos conforme quadro 4 e aplicando o Modelo de Classificação Logístico obteve-se, através da matriz de confusão, os índices de precisão apresentados no quadro 6 assim como a curva ROC do modelo produzido e apresentada na figura 2.

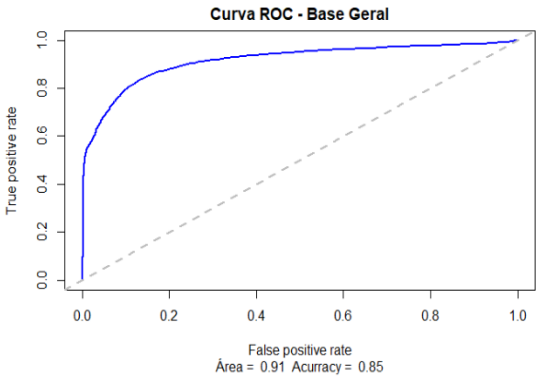



**Figura 2 -** Curva ROC do modelo gerado a partir da base geral

**Quadro 5 -** Variáveis e seus índices de significância, selecionadas através da Regressão *Stepwise* aplicada a Base Geral

| | Estimate | Std.Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 0.0129876 | 0.0129876 | 78.122 | < 2e-16 *** |
| VAR01 | 0.0011975 | 0.0003955 | 3.028 | 0.002487 ** |
| VAR06 | 0.0484049 | 0.0306022 | 1.582 | 0.113841 |
| VAR08 | 0.0227512 | 0.0062065 | 3.666 | 0.000252 *** |
| VAR09 | 0.0148059 | 0.0047020 | 3.149 | 0.001660 ** |
| VAR12 | -0.0255182 | 0.0127252 | -2.005 | 0.045042 * |
| VAR14 | 0.0614615 | 0.0073499 | 8.362 | < 2e-16 *** |
| VAR16 | -0.1459795 | 0.0209859 | -6.956 | 4.51e-12 *** |
| VAR17 | 0.0029227 | 0.0005390 | 5.422 | 6.48e-08 *** |
| VAR18 | 0.0021749 | 0.0005241 | 4.150 | 3.45e-05 *** |
| VAR21 | 0.0286768 | 0.0054393 | 5.272 | 1.47e-07 *** |
| VAR23 | -0.0135823 | 0.0058131 | -2.336 | 0.019549 * |
| VAR25 | -0.0007355 | 0.0001780 | -4.132 | 3.72e-05 *** |
| VAR31 | -0.0029649 | 0.0006821 | -4.347 | 1.44e-05 *** |
| VAR32 | -0.0010347 | 0.0003187 | -3.247 | 0.001182 ** |
| VAR33 | -0.4312691 | 0.0137166 | -31.441 | < 2e-16 *** |
| VAR34 | -0.0374658 | 0.0051717 | -7.244 | 5.85e-13 *** |
| VAR35 | 0.0522678 | 0.0260286 | 2.008 | 0.044747 * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Quadro 6 -** Índices de precisão do modelo Logístico obtido através da base geral

| Índices | Base Geral |
|---|---|
| **Accuracy** | 0.85 |
| **Kappa** | 0.70 |
| **Sensitivity** | 0.83 |
| **Specificity** | 0.84 |
| **Precision** | 0.87 |
| **Prevalence** | 0.50 |
| **Balanced_Accuracy** | 0.85 |

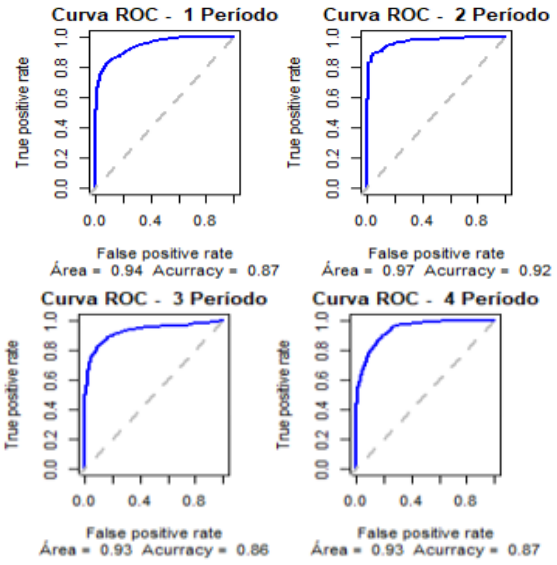
## 7.2 Cenário "Modelo por Período"

Neste modelo de dados foi utilizado a base geral recortada por período sem discriminação de Curso, em cada um dos oito recortes gerados foi aplicado o modelo de classificação Logístico e através da matriz de confusão foram obtidos os índices de precisão descritos no quadro 7 e a curva ROC para cada Período com as respectivas áreas abaixo da curva e *accuracy* do modelo, conforme apresentado na figura 3.

**Quadro 7 -** Índices de precisão do modelo Logístico obtido através dos recortes por período

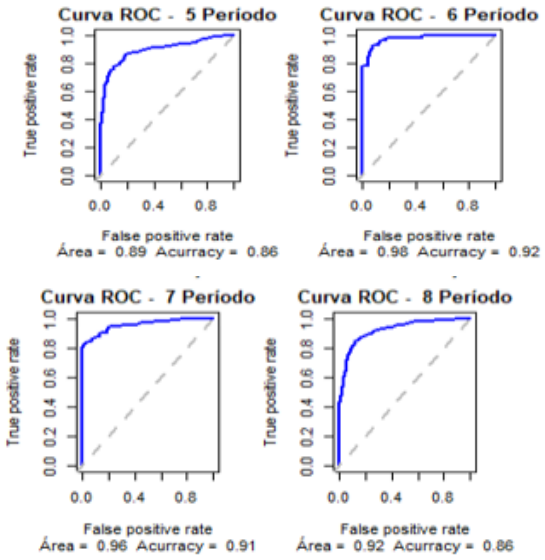
| Índices | Períodos | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Accuracy** | 0.87 | 0.92 | 0.86 | 0.87 | 0.86 | 0.92 | 0.91 | 0.86 |
| **Kappa** | 0.74 | 0.84 | 0.72 | 0.71 | 0.69 | 0.83 | 0.81 | 0.71 |
| **Sensitivity** | 0.79 | 0.87 | 0.84 | 0.93 | 0.72 | 0.88 | 0.84 | 0.86 |
| **Specificity** | 0.85 | 0.89 | 0.87 | 0.85 | 0.86 | 0.91 | 0.87 | 0.89 |
| **Precision** | 0.91 | 0.96 | 0.86 | 0.87 | 0.87 | 0.93 | 0.96 | 0.82 |
| **Prevalence** | 0.45 | 0.47 | 0.46 | 0.64 | 0.36 | 0.45 | 0.48 | 0.43 |
| **Balanced_ Accuracy** | 0.86 | 0.92 | 0.86 | 0.84 | 0.83 | 0.91 | 0.91 | 0.86 |

**Figura 3:** Curvas ROC dos modelos gerados a partir dos recortes por Período





http://dx.doi.org/10.25286/repa.v3i3.954




## 7.3 Cenário "Modelo por Curso"

Neste modelo de dados foi utilizado a base geral recortada por Curso sem discriminação de Período, em cada um dos quatro recortes gerados foi aplicado o modelo de Regressão Logístico e através da matriz de confusão foram obtidos os índices de precisão descritos no quadro 8 e a curva ROC para cada Período com as respectivas áreas abaixo da curva e *accuracy* do modelo, conforme apresentado na figura 4.

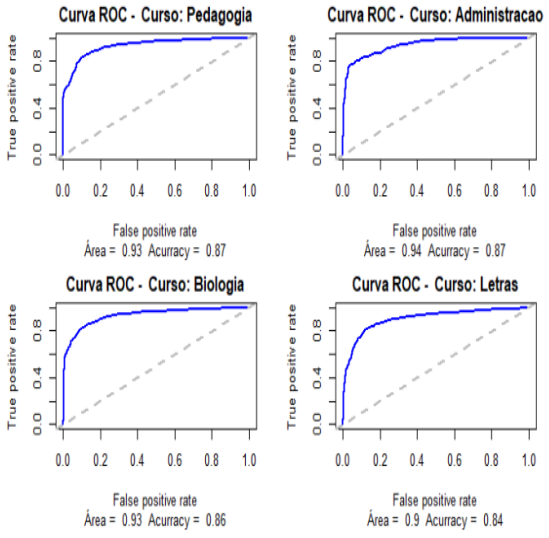



**Figura 4 -** Curvas ROC dos modelos gerados a partir dos recortes por Curso

**Quadro 8**: Índices de precisão dos modelos gerados a partir dos recortes por Curso

| Índices | Pedag | Administ | Biologia | Letras |
|---|---|---|---|---|
| **Accuracy** | 0.87 | 0.87 | 0.86 | 0.84 |
| **Kappa** | 0.73 | 0.73 | 0.72 | 0.69 |
| **Sensitivity** | 0.86 | 0.81 | 0.86 | 0.83 |
| **Specificity** | 0.86 | 0.86 | 0.86 | 0.83 |
| **Precision** | 0.87 | 0.88 | 0.87 | 0.86 |
| **Prevalence** | 0.50 | 0.44 | 0.50 | 0.52 |
| **Balanced_ Accuracy** | 0.87 | 0.86 | 0.86 | 0.84 |

## 8 Avaliação dos Modelos

Neste trabalho foi aplicado análise de Regressão Logística na geração de um modelo para previsão de reprovação de estudantes através dos números de interações dos mesmos com o ambiente virtual de ensino.

Através do modelo de dados Base Geral, sem recortes, o Modelo de Regressão Logística gerado apresentou um índice de *accuracy* de 85%, em busca de melhorar a acurácia deste modelo a base de dados foi recortada por Curso e Período, aplicada a Regressão Logística em cada recorte e calculada a média dos índices de desempenho dos modelos gerados através dos referidos recortes conforme quadro 9, gerou-se dois vetores com as medias dos índices de desempenho em cada recorte e outro vetor com os índices do modelo de dados sem cortes, nestes três vetores foi aplicada análise de variância (ANOVA) que resultou em um valor p associado a estatística t igual a 0,955 conforme demonstrado no quadro 10, portanto pode-se afirmar que não há diferença entre as médias dos índices de desempenho dos modelos aplicados a base geral e os recortes, sendo assim pressupõe-se que os modelos de regressão Logística obtidos a partir da base de dados geral e seus recortes tem índices de desempenho semelhantes o que podemos constatar visualmente na curva ROC dos referidos modelos.

**Quadro 9** - Índices de *accuracy* do modelo gerados a partir da Base geral e média dos índices gerados a partir dos recortes por Período e Curso

| Índice | Base geral | Índice médio por recorte | |
|---|---|---|---|
| | | Curso | Período |
| **Accuracy** | 0.85 | 0.86 | 0.88 |
| **Kappa** | 0.70 | 0.72 | 0.76 |
| **Sensitivity** | 0.83 | 0.84 | 0.84 |
| **Specificity** | 0.84 | 0.85 | 0.87 |
| **Precision** | 0.87 | 0.87 | 0.90 |
| **Prevalence** | 0.50 | 0.49 | 0.47 |
| **Balanced_ Accuracy** | 0.85 | 0.86 | 0.88 |

**Quadro 10 –** A nova aplicada aos índices dos modelo apresentados na Quadro 2.

| | Df | SumSq | Mean Sq | Fvalue | Pr(>F) |
|---|---|---|---|---|---|
| **Bases** | 2 | 0,0016 | 0.000817 | 0.046 | 0.955 |
| **Residuals** | 21 | 0.3707 | 0.017651 | | |

De acordo com as análises realizadas, o Modelo de Classificação Logístico para previsão de reprovação do aluno pode ser estimado através do modelo de dados generalizado através da base geral sem recortes com índice de desempenho mediano de 0.835, através dos modelos de dados gerados a partir dos recortes por Curso os quais apresentaram índice de desempenho mediano de 0.845 ou através dos modelos de dados gerados a partir dos recortes por Período que apresentaram índice de desempenho mediano de 0,855 conforme o quadro 10.

**Quadro 10 -** Sumário dos índices de desempenho por modelo e dados.

| Índice | Base Geral | Curso | Período |
|---|---|---|---|
| **Min.** | 0.5000 | 0.4900 | 0.4700 |
| **1st Qu.** | 0.7975 | 0.8100 | 0.8200 |
| **Median** | 0.8350 | 0.8450 | 0.8550 |
| **Mean** | 0.7837 | 0.7913 | 0.8037 |
| **3rd Qu.** | 0.8500 | 0.8600 | 0.8725 |
| **Max.** | 0.8700 | 0.8700 | 0.9000 |

Com o *BoxPlot* comparativo apresentado na figura 5 podemos verificar que os índices de desempenho possuem distribuição assimétrica, variabilidades semelhantes e que os modelos gerados através dos recortes por Curso apresentam índices de desempenho mediano mais próximo do modelo gerado a partir da Base Geral e que apesar do modelo gerado a partir dos recortes por Período apresentar índices de precisão levemente superior aos demais modelos, estatisticamente a ANOVA comprova a igual capacidade preditiva dos três modelos.

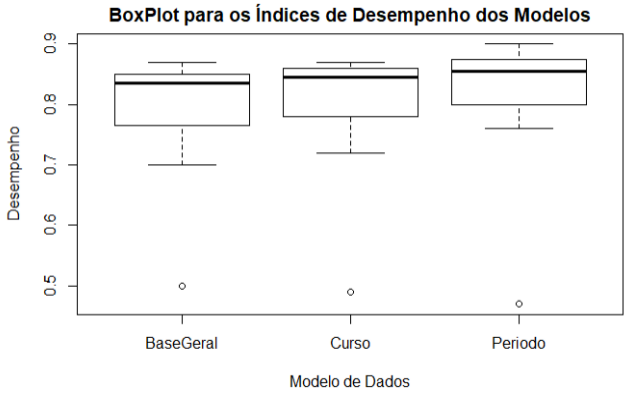



**Figura 5 -** Gráfico box-plot, índices desempenho dos modelos de classificação

## 9 Conclusões

A consistência dos dados através de forma padronizada e estruturada é fundamental para o sucesso das pesquisas e implementação de modelos de classificação e previsão.

Neste trabalho, através de algumas análises descritivas como visto em outros trabalhos, constatou-se que o desempenho do aluno está diretamente associado ao total de interações pois os alunos que tiveram desempenho superior também realizaram uma número superior de acessos ao ambiente de ensino.

Como foi demonstrado, os modelos de previsão obtidos através da Regressão Logística realizada através dos três modelos de dados obtiveram excelentes índices de desempenho possibilitando a aplicação de qualquer um dos modelos na previsão de reprovação do aluno, portanto tomando como referência a Base Geral sem recorte, conforme variáveis selecionadas e apresentadas no quadro 4, com nível de confiança de 95% e significância de 5% os meios de interação que mais contribuem com a previsão de


http://dx.doi.org/10.25286/repa.v3i3.954

reprovação do aluno são representados através da variáveis VAR12, VAR23 e VAR35, com nível de confiança de 99% e nível de significância de 1% os meios de interação que mais contribuem são representados através das variáveis VAR01, VAR09 e VAR32. Assim como na figura 4, seria possível verificar as variáveis que mais contribuem com a previsão de reprovação nos demais modelos de dados.

Apesar das análises realizadas, não podemos desconsiderar que o modelo de previsão pode cometer o grave erro de apresentar elevada taxa de falso positivo em que o algoritmo classifica um aluno no grupo de aprovação enquanto o mesmo encaminha-se para a reprovação ou um erro menos grave, falso negativo, em que o aluno erroneamente é classificado no grupo de reprovados, erros esses que podem comprometer o modelo de classificação.

## Agradecimentos

Os autores agradecem a colaboração do Núcleo de Educação a Distância - NEAD/UPE pelo fornecimento da base de dados. Os autores também agradecem ao Grupo de Pesquisa em Ciência de Dados Educacionais (CiDE/UFRPE).

## Referências

**[1]** CECHINEL, Cristian; ARAUJO, Ricardo Matsumura; DETONI, Douglas. Modelling and Prediction of Distance Learning Students Failure by using the Count of Interactions. **Brazilian Journal of Computers in Education**, v. 23, n. 03, p. 1, 2015.

**[2]** ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601-618, 2010.

**[3]** BROWN, Malcoln. Learning Analytics: Moving from Concept to Practice, **EDUCAUSE Learning Initiative Brief**, v. 7, 2012.

**[4]** MURRAY, Meg et al. Student interaction with content in online and hybrid courses: Leading horses to the proverbial water. In: INFORMING SCIENCE AND INFORMATION TECHNOLOGY EDUCATION CONFERENCE, 2013, Porto. **Proceedings**…Porto: Informing Science Institute, 2013. p. 99-115.

**[5]** DICKSON, W. Patrick. Toward a deeper understanding of student performance in virtual high school courses: Using quantitative analyses and data visualization to inform decision making. **A synthesis of new research in K–12 online learning**, p. 21-23, 2005.

**[6]** MANHÃES, Laci Mary Barbosa et a. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: *WORKSHOP* DE INFORMÁTICA NA ESCOLA, SIMPÓSIO BRASILEIRO DE INFORMATICA NA EDUCAÇÃO, 22., 2011, Aracajú. Anais..., Aracaju, 2011.

**[7]** BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011.

**[8]** BATISTA, Antonio Sarmento. **Regressão Logística**: uma introdução ao modelo estatístico. Porto: Vida Econômica, 2015.