

# Predição de reprovação na educação a distância: um estudo comparativo

Patrícia Takaki Neves<sup>1</sup>, Bruno Elias Pentead<sup>2</sup>, Seiji Isotani<sup>2</sup>

<sup>1</sup>Docente do Depto. de Ciências da Computação (UNIMONTES), Mestre em Ciência da Computação (UNICAMP), Doutoranda em Ciência da Informação (UFSC) e Pós Graduanda em Computação Aplicada à Educação (ICMC/USP)

<sup>2</sup> Universidade de São Paulo (ICMC/USP)

## INTRODUÇÃO

- **TEMA:** Mineração de Dados Educacionais (MDE).
- **PROBLEMA:** Predição da reprovação de alunos da EaD.
- **JUSTIFICATIVA:** Crescimento da EaD / Disponibilidade dos dados nos AVA / Desperdício de recursos com reprovações e evasões / Apoio à ação docente / Uso das técnicas da MDE.

## OBJETIVOS

- Realizar um estudo comparativo de diferentes modelos preditivos de reprovação.
- Utilizar 6 algoritmos, 2 conjuntos de dados normalizados e 3 opções de balanceamento de dados.
- Comparar 5 métricas de desempenho: acurácia geral, sensibilidade, especificidade, índice *kappa* e *g-means*.

## MATERIAIS E MÉTODOS

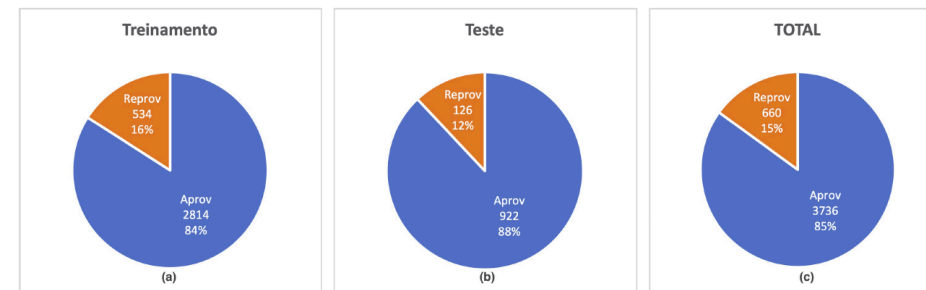
- **MATERIAIS:** software WEKA e 4396 conjuntos de dados do Moodle (relatórios de 21 disciplinas cursadas por 254 alunos que ingressaram no Curso de Pedagogia da UAB/Unimontes).
- **MÉTODOS:** pré-processamento, divisão dos dados de 3 e 5 semanas (de um total de 6); divisão dos dados para treinamento e testes; execução dos algoritmos *Naive Bayes*, *Logistic Regression*, *Support Vector Machine*, *Decision Table*, *J48 (Decision Tree)* e *OneR*; e por fim comparação e discussão das métricas.

## RESULTADOS

- Dados desbalanceados (aprovados e reprovados)

Tabela 1. Aprovações e reprovações

STATUS	Treinamento	Teste	TOTAL
Aprov	2814	922	3736
Reprov	534	126	660
<b>TOTAL</b>	<b>3348</b>	<b>1048</b>	<b>4396</b>



- Melhores resultados: SVM (com SMOTE em acurácia e índice *kappa* e sem balanceamento em especificidade) e *Decision Table* (com *ClassBalancer*).
- Filtro de balanceamento melhora sensibilidade e *g-means*.
- Dados de 3 semanas tiveram bom desempenho.
- É preciso considerar os objetivos e contextos educacionais envolvidos para interpretar as métricas e modelos.
- No futuro: consultar especialistas, usar *f-measure*, aplicar testes com todos os cursos, acrescentar dados demográficos e cadastrais e incorporar no Moodle.

Tabela 2. Resultados com dados de 3 semanas

Algoritmo	Filtro para balanceamento	dataset 3 semanas					
		Acurácia	Sensibilidade	Especificidade	G-means	Kappa	
Naive Bayes	sem	<b>88,8%</b>	60,3%	<b>92,7%</b>	74,8	0,501	
	com SMOTE	88,7%	<b>61,9%</b>	92,4%	<b>75,6</b>	<b>0,505</b>	
Logistic	com ClassBalancer	88,5%	<b>61,9%</b>	92,2%	75,5	0,500	
	sem	<b>92,4%</b>	55,6%	<b>97,4%</b>	73,6	<b>0,595</b>	
SMO (SVM)	com SMOTE	88,7%	61,9%	92,4%	<b>75,6</b>	0,505	
	com ClassBalancer	82,7%	<b>66,7%</b>	84,9%	75,3	0,387	
Decision Table	sem	92,6%	45,2%	<b>99,0%</b>	66,9	0,557	
	com SMOTE	<b>93,1%</b>	65,1%	97,0%	<b>79,5</b>	<b>0,656</b>	
OneR	com ClassBalancer	81,9%	<b>66,7%</b>	83,9%	74,8	0,371	
	sem	90,5%	52,4%	<b>95,7%</b>	70,8	0,516	
J48	com SMOTE	83,4%	61,9%	86,3%	73,1	0,381	
	com ClassBalancer	<b>90,9%</b>	<b>70,6%</b>	93,7%	<b>81,3</b>	<b>0,600</b>	
OneR	sem	<b>92,8%</b>	<b>59,5%</b>	<b>97,4%</b>	<b>76,1</b>	<b>0,627</b>	
	com SMOTE	90,2%	48,4%	95,9%	68,1	0,488	
J48	com ClassBalancer	sem	<b>92,8%</b>	52,4%	<b>97,7%</b>	71,6	<b>0,578</b>
	com SMOTE	77,1%	65,9%	78,6%	72,0	0,291	
	com ClassBalancer	76,3%	<b>80,2%</b>	75,8%	<b>78,0</b>	<b>0,334</b>	

Figura 7. Análise temporal do SVM com SMOTE com dados de 1 a 5 semanas

