12º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2021

Modelos de Machine Learning para previsão de evasão em um curso de Licenciatura em Matemática do IFSP

Vanessa Cassiana da Silva[1], Michael Macedo Diniz[2]

[1] Licenciatura em Matemática, Bolsista PIBIC SP, IFSP, Câmpus São José dos Campos, vanessa.cassiana@aluno.ifsp.edu.br.
[2] Docente na área de Matemática no IFSP, Câmpus São José dos Campos, michael.diniz@ifsp.edu.br.
Área de conhecimento (Tabela CNPq): 1.01.00.00-8 Matemática

RESUMO: A evasão em cursos de ensino superior é um problema que se destaca devido às perdas econômicas, sociais e acadêmicas em instituições de ensino superior do Brasil e do mundo. No Brasil, há três modalidades principais para o abandono do curso, sendo elas a econômica, a institucional e a vocacional. Ainda neste contexto, a Mineração de Dados Educacionais (EDM) é uma ferramenta cada vez mais utilizada para a resolução de problemas no âmbito educacional. Neste trabalho utilizamos a Regressão Logística (RL), o K-Nearest Neighbors (KNN) e o Random Forest, para prever a evasão de alunos no 1º semestre do curso de Licenciatura em Matemática do IFSP-SJC. Para treinar os modelos foram utilizados bancos de dados contendo os dados de todos os alunos que realizaram a matrícula no curso entre os anos de 2016 a 2019. Os resultados foram promissores utilizando o KNN e RL, porém a partir do KNN obteve-se uma acurácia superior a 80%, sendo, então, o melhor modelo para esse tipo de problema.

PALAVRAS-CHAVE: Regressão Logística; KNN; Random Forest; Educational Data Mining; evasão.

Machine Learning models to predict student drop out in a Mathematics' course of the IFSP.

ABSTRACT: College dropout is a problem that stands out due to economics, social and academic losses in higher education institutions in Brazil and the world. In Brazil, there are three main factors for college dropout being the economic, the institutional and the vocational. In this context, the Educational Data Mining (EDM) is an increasingly used tool for solving educational problems. In this paper, Logistic Regression (LR), K-Nearest Neighbors (KNN) and Random Forest, Machine Learning models, were used to predict student dropout in the 1st semester of the Mathematics Degree course at IFSP-SJC. For the model, databases containing the data of all students who enrolled in the course between the years 2016 to 2019 were used. The results were promising using the KNN and LR, furthermore from the KNN a higher accuracy was obtained. at 80%, thus being the best model for this kind of problem.

KEYWORDS: Logistic Regression; KNN; Random Forest; Educational Data Mining; dropout.

INTRODUÇÃO

A evasão estudantil no ensino superior é um problema generalizado que afeta o resultado dos sistemas educacionais. De acordo com Silva Filho (2007), a evasão implica em desperdícios sociais,

acadêmicos e econômicos, tanto no setor público, quanto no privado. De modo geral, recursos não aproveitados de professores, funcionários, equipamentos e espaço físico.

A evasão em Instituições de Ensino Superior do Brasil (IES) pode ser explicada por três causas principais de acordo com Barroso e Falcão (2004), são elas: a evasão econômica, a evasão vocacional e a evasão institucional, sendo esta podendo ser evitada com planejamento docente.

Tendo em vista que a evasão é um problema que aflige os alunos, os docentes e as IES, busca-se tomar medidas preventivas a fim de diminuí-la e com o objetivo de prever os alunos mais propensos a evasão para uma intervenção institucional ou docente, é possível criar modelos matemáticos a partir de técnicas de Data Mining voltadas para Educação (EDM).

No artigo iremos apresentar dois modelos para prever cada uma das seguintes situações: 1) se o aluno irá evadir no primeiro semestre do curso de Licenciatura em Matemática, e 2) se o aluno irá evadir no segundo semestre do curso. Para tanto, foram testados 3 modelos de classificação: Regressão Logística, KNN e Random Forest.

MATERIAL E MÉTODOS

Foram estruturados dois bancos de dados, o primeiro banco refere-se ao primeiro objetivo que consiste em prever os alunos que irão evadir no primeiro semestre. Este banco contém dados referente aos alunos que estão matriculados em matérias do 1º semestre dos anos de 2016, 2017, 2018 e 2019. As variáveis que o compõe são: “Nome”, “RA”, “Ano de ingresso”, “Opção de curso”, “Nota de Linguagens no ENEM”, “Nota de Ciências Humanas no ENEM”, “Nota de Ciências da Natureza no ENEM”, “Nota de Matemática no ENEM”, “Nota de Redação no ENEM”, “Nota do ENEM”, “Cidade de origem”, “Meio de transporte”, “Tipo de escola”, “Tipo de ingresso no IFSP”, “Renda familiar per capita”, “Renda familiar bruta”, “Sexo” e “Evasão”. Porém, as variáveis "Nome" e "RA" foram retiradas por serem da classe *string* e não influenciarem no resultado. As variáveis “Renda familiar per capita” e “Renda familiar bruta” também foram retiradas por terem mais da metade dos dados faltantes.

O segundo banco contém dados dos alunos matriculados em matérias do segundo semestre dos anos de 2016, 2017, 2018 e 2019. As variáveis que o compõe são as mesmas do primeiro semestre, além de: “Quantidade de reprovações no 1º semestre”, “IRA em exatas no 1º semestre”, “IRA em humanas no 1º semestre”, “Quantidade de reprovações por falta no 1º semestre”, “Média de frequência das disciplinas que cursou no 1º semestre”, “Média final em Leitura Interpretação e Produção Textual (LIP)”, “Média final em História da Educação (HED)”, “Média final em Estatística Básica (EST)”, “Média final em Fundamentos de Geometria Analítica 1 (FGA 1)", “Média final em Fundamentos de Matemática Elementar 1 (FM1)”, e “Média final em Geometria Euclidiana 1 (GE1)”. Neste banco, as variáveis “Nome”, “RA”, “Renda familiar per capita” e “Renda familiar bruta”, também foram retiradas.

No Python, foi utilizada a biblioteca imbalanced-learn para balancear os dados utilizando a técnica SMOTE (Laureano et al. 2020), que gera dados sintéticos da classe minoritária a partir de vizinhos. Os dados foram divididos e treinados a partir do SciKit-Learn para todos os modelos. Em relação aos bancos, foram utilizados todos os classificadores com os parâmetros ajustados pelo Grid Search em cada um. Para cada um dos modelos foram feitos os relatórios de classificação que retorna a acurácia, a precisão e o recall de cada classificador. A partir da matriz de confusão descrita na Tabela 1, as métricas de avaliação são definidas.

TABELA 1. Matriz de confusão. A diagonal principal representa todos os dados classificados corretamente.

		CLASSIFICADO	
		SIM	NÃO
REAL	SIM	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	NÃO	Falso Positivo (FP)	Verdadeiro Negativo(VN)

A acurácia é uma métrica que indica a performance geral do modelo, ou seja, quantas classificações foram feitas corretamente. A precisão refere-se à quantidade de classificações da classe positiva que o modelo fez corretamente, por fim, o recall refere-se à quantidade de classificações positivas em relação à amostra real. As suas fórmulas estão apresentadas abaixo.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$Precisão = \frac{VP}{VP + FP} \quad (2)$$

$$Recall = \frac{VP}{VP + FN} \quad (3)$$

RESULTADOS E DISCUSSÃO

Banco 1

No banco dos alunos evadidos no primeiro semestre, o primeiro modelo testado foi o de Regressão Logística. Com o Grid Search, obteve-se os parâmetros a serem ajustados que estão descritos na Tabela 2, assim como os valores do parâmetro e o melhor valor para o modelo.

TABELA 2. Descrição dos parâmetros de Regressão Logística a serem ajustados com o Grid Search, seus valores e os resultados obtidos.

Parâmetros	Valores	Melhor
C	$[10^{-4}, 10^{4}]$	1438.44988828766
intercept_scaling	{1, 2, …, 39, 40}	5
multi_class	‘auto’, ‘ovr’, ‘multinomial’	‘ovr’
penalty	‘l1’, ‘l2’, elasticnet’	‘l2’
solver	‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’	‘liblinear’

Com o *Cross Validation* de 10 *folds*, a acurácia, a precisão e o recall obtidos foram 66% para as três métricas.

Quanto ao KNN, com o Grid Search, os parâmetros testados e obtidos, assim como seus valores, estão descritos na Tabela 3, que em comparação com o modelo de Regressão Logística, tem uma menor quantidade. Nenhum dos parâmetros permaneceu o mesmo.

TABELA 3. Descrição dos parâmetros do KNN a serem ajustados com o Grid Search, seus valores e o resultado obtido.

Parâmetros	Valores	Melhor
n_neighbors	{1,, 2, …, 39, 40}	4
weights	‘distance’, ‘uniform’	‘distance’
p	{1, 2, …, 8, 9}	1

Utilizando o *Cross Validation* para o KNN, a acurácia, precisão e recall obtidos foram de 76%, um resultado superior ao modelo de RL.

Por fim, o último modelo a ser testado foi o Random Forest, com o Grid Search, os parâmetros testados e obtidos estão descritos na Tabela 4.

TABELA 4. Descrição dos parâmetros do Random Forest a serem ajustados com o Grid Search, seus valores e o resultado obtido.

Parâmetros	Valores	Melhor
n_estimators	{100, 101, …, 9998, 9999}	1948
criterion	‘gini’, ‘entropy’	‘entropy’
max_features	‘auto, ‘sqrt’, ‘log2’	‘sqrt’
n_jobs	[1,40]	22

Com o *Cross Validation*, a acurácia obtida foi de 60%, a precisão foi de 59% e o recall foi de 69%.

Além disso, é possível fazer uma correlação entre a variável ‘Evasão’ e o restante das variáveis utilizadas para entender melhor os resultados.

TABELA 5. Correlação entre a variável evasão e as demais variáveis. A correlação estabelece um valor do quanto as variáveis estão relacionadas, quanto maior o valor, maior a relação.

Variável	Correlação
Ano de ingresso	0.054295
Idade no 1º semestre	0.034927
Opção de curso	-0.288712
Nota LIN	0.060715
Nota CH	-0.005920
Nota CN	-0.064919
Nota MAT	-0.162911
Nota RED	-0.009541
Nota ENEM	-0.127758
Cidade de Origem	-0.066504
Meio de transporte	0.013204
Escola	0.081347
Tipo de ingresso	0.115796
Sexo	-0.174716
Evasão	1

Em Brito et al. (2014), sobre desempenho de alunos e evasão, as notas do ENEM e as notas das matérias do primeiro semestres, foram as variáveis mais significativas para a classificação. No banco um, que contém as notas do ENEM, a correlação entre essas notas e a variável evasão não foi alta, o que pode ter influenciado no resultado, delimitando-o a 76% de acurácia. De modo geral, no primeiro banco, o modelo de maior sucesso foi o KNN.

Banco 2

No segundo banco, referente aos alunos que evadiram depois do primeiro semestre, os parâmetros dos modelos e as classes a serem otimizadas permaneceram os mesmos.

Com a Regressão Logística e ajuste de parâmetros, sendo C = 29.76351441631312, intercept_scaling = 23, multi_class: ‘over’, penalty: ‘l2’ e solver: ‘saga’, e com o Cross Validation 10 folds, obteve-se uma acurácia de 85%, precisão de 79% e recall de 96%.

Em relação ao KNN, com a otimização obtivemos os parâmetros n_neighbors = 1, p = 1 e weights:’distance’, e com o Cross Validation 10 folds, obteve-se uma acurácia de 89%, precisão de 82% e recall de 100%.

Por fim, o Random Forest foi testado, a acurácia obtida com os parâmetros padrões foi de 57%, enquanto a precisão foi de 65% e o recall de 47%, considerando os alunos que evadiram.

Assim como no banco 1, uma análise da correlação também foi realizada e está descrita na Tabela 6 abaixo. Neste banco é possível notar que há uma correlação maior entre as demais variáveis e a variável de evasão.

TABELA 6. Correlação entre a variável evasão e as demais variáveis.

Variável	Correlação
Ano de ingresso	0.067200
Reprovações no 1º semestre	-0.445399
IRA em exatas	0.509428
IRA em humanas	0.456340
Reprovações por falta	-0.308043
Média da frequência	0.483354
Idade	-0.111684
MF_LIP	0.189922
MF_HED	0.388469
MF_EST1M1	0.329582
MF_FGAM 1	0.255088
MF_FM1M1	0.305895
MF_GE1ME	0.176016
Opção de curso	-0.033952

Nota LIN	0.138272
Nota CH	0.140113
Nota CN	0.042933
Nota MAT	0.241754
Nota RED	0.096213
Nota ENEM	0.212241
Cidade de Origem	-0.058999
Meio de transporte	-0.166932
Escola	-0.027488
Tipo de ingresso	-0.041046
Sexo	0.023395
Evasão	1

Como descrito no banco um, de acordo com Brito et al. (2014), as variáveis que mais influenciam são as notas do ENEM e as notas das matérias do primeiro semestre. No banco dois, isso se mostrou mais efetivo, uma vez que os resultados superaram os 79% de acurácia obtida no banco um, além disso, a maior correlação foi com as variáveis "IRA em exatas", "IRA em humanas" e "Média da frequência", o que mostra que as notas e as presenças são de maior importância para a classificação.

CONCLUSÕES

No primeiro banco os resultados alcançaram 76% de acurácia, enquanto no segundo banco os resultados foram mais promissores, alcançando 100% de precisão, mostrando, dessa forma, que o modelo classificou corretamente os alunos evadidos. Em ambos os bancos os melhores resultados foram obtidos a partir do KNN.

Ainda assim, o segundo banco obteve os melhores resultados, acredita-se que seja pela maior correlação entre os dados de entrada e a variável de evasão, desse modo, quanto maior a correlação entre as variáveis melhor serão os resultados dos modelos.

Em relação a bibliografia utilizada no projeto, os resultados se mostraram satisfatórios e até maiores do que a média. Badr et al. (2016), que busca prever a performance de estudantes de exatas a partir de matérias do primeiro semestre a partir de modelos de classificação, obteve 52.94% de acurácia. Viloria et al. (2019), que buscou analisar os casos de evasão entre estudantes do ensino superior a partir de classificação, obteve 77% de acurácia. E, finalmente, Brito et al. (2014), que a partir de dados do ENEM classificaram o desempenho dos alunos do primeiro semestre, os resultados foram de 74.67%.

AGRADECIMENTOS

Agradeço à CNPq, processo nº 23305.006663.2020-61, pela oportunidade e auxílio financeiro para realizar essa Iniciação Científica.

REFERÊNCIAS

Barroso, M. F. e Falcão, E. B. Evasão universitária: o caso do Instituto de Física da UFRJ. IX Encontro Nacional de Pesquisa em Ensino de Física. 2004.

Laureano, L. B., Sison, A. M., Medina, R. P. Affinity Propagation SMOTE approach for Imbalanced dataset used in Predicting Student at Risk of Low Performance. International Journal of Advanced Trends in Computer Science and Engineering, v.9, n. 4, pp. 5066 - 5070. 2020

Rodrigues, R. L., Medeiros, F. P. A. de, e Gomes, A. S. Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. DOI: 10.5753/CBIE.SBIE.2013.607.

Silva Filho, R. L. L., Montejunas, P. R., Hipólito, O. e Lobo, M. B. C. M. A evasão no ensino superior brasileiro. Cadernos de Pesquisa, v. 37, n. 132, pp. 641- 659. 2007