APRENDIZADO DE MÁQUINA COM DADOS SOCIOECONÔMICOS PARA APOIAR O CONTROLE DE EVASÃO

Gabriélle Helpis dos Santos[1], Gilson Saturnino dos Santos[1]

[1]Instituto Federal do Mato Grosso do Sul– Coxim-MS

gabrielle.helpis@gmail.com, gilson.santos@ifms.edu.br

Resumo

As Instituições de Ensino, inclusive as de nível superior, enfrentam dificuldades em cessar a evasão dos mais variados cursos. Alguns instituições realizam ações com objetivo de aumentar a permanência dos estudantes. A identificação antecipada de estudantes com possibilidade de desistência poderia contribuir para uma melhor atuação da instituição em relação a redução da evasão. Sendo assim, o presente trabalho tem por objetivo utilizar o Aprendizado de Máquina para apoiar o controle de evasão. Foram gerados 9 classificadores baseados em respostas de questionários socioeconômicos, de estudantes de cursos superiores de 8 campi do Instituto Federal de Mato Grosso do Sul (IFMS). Os classificadores foram gerados e avaliados utilizando o Weka e quatro algoritmos de classificação: *Naive Bayes*, *Multilayer Perceptron*, *SMO* e *J48*. Espera-se que os resultados obtidos possam auxiliar na análise de Evasão, possibilitando a utilização dos modelos obtidos para criar sistemas que contribuam com a redução da evasão de cada Campus do IFMS.

Palavras-chave: Evasão Escolar, Aprendizado de Máquina, Mineração de Dados.

Introdução

A evasão nas Instituições de Ensino Superior (IES), é um problema que vem se agravando mais a cada ano. No entanto poucas IES contam com um programa especializado de combate à evasão, podendo ser um dos motivos do problema não estar regredindo (CARDOSO; LUDOVICO, 2017). A identificação antecipada de estudantes que podem evadir poderia contribuir com a eficiência destes programas.

Com o avanço da tecnologia, é cada vez mais fácil acessar e guardar todo tipo de informação. Diversas IES têm um vasto banco de dados com informações relevantes sobre seus alunos. Deste modo é possível utilizar a mineração de dados educacional para responder questões da Educação, como a Evasão e os motivos que levam a ela (RIGO, BARBOSA; CAMBRUZZI, 2012).

Uma possível abordagem para identificar com antecedência estudantes com tendência a evasão seria a utilização de conceitos de Aprendizado de Máquina (AM). Um sistema de aprendizado é um software que toma decisões baseado em aprendizado realizado a partir de soluções corretas de problemas anteriores. Estes sistemas utilizam conceitos e algoritmos da área da computação chamada de AM, uma subárea da Inteligência Artificial (AMARAL, 2016).

O objetivo do trabalho foi utilizar o Aprendizado de Máquina, com a ferramenta *open-source* Weka, que é uma ferramenta e biblioteca em Java (WITTEN E FRANK, 2011), para criar e avaliar modelos de classificação que possam identificar estudantes que irão evadir de cursos superiores, apoiando assim o controle de Evasão do Instituto Federal do Mato Grosso do Sul (IFMS).

Metodologia

Inicialmente foi realizada uma pesquisa sobre evasão no ensino superior e aprendizado de máquina no apoio ao combate à evasão. Em seguida, foi realizado um estudo sobre a ferramenta *open-source* de Aprendizado de Máquina Weka e suas funcionalidades.

Posteriormente, foi realizado um estudo detalhado sobre o banco de dados socioeconômicos e da situação (Evadido e Não Evadido) dos estudantes do IFMS, de cursos superiores, a fim de gerar um modelo entidade-relacionamento (MER) para melhor compreensão da distribuição das tabelas e seus relacionamentos. O banco de dados conta com 24 tabelas, sendo que a tabela que possui mais campos é a tabela *dados*, com 44 campos, enquanto as tabelas *grupos*, *locais*, *sistemas* e *versoes*, contam com apenas duas campos.

Após a criação do MER, foram realizadas diversas subconsultas no banco de dados que geraram 9 tabelas com dados específicos das respostas dos alunos em relação ao questionário socioeconômico, sendo 8 dos Campi dos IFMS, e uma onde não há distinção entre os Campi. Na tabela 1 pode-se verificar um total de 2698 respostas de questionário socioeconômicos.

Campus	Quantidade de Registros
Aquidauana	224
Campo Grande	365
Corumbá	448
Coxim	429
Naviraí	77
Nova Andradina	331
Ponta Porã	536
Três Lagoas	288
IFMS (Geral)	2698

Tabela 1. **Quantidade de Respostas do Questionário Socioeconômico**. Fonte: Autoria Própria.

Por último, a partir do banco de dados do IFMS foi utilizado o Weka para gerar e avaliar os classificadores, com quatro diferentes algoritmos: Naive Bayes, Rede Neural Artificial

Multilayer Perceptron, Máquina de Vetor de Suporte SMO e a Árvore de Decisão J48. Todos utilizando a metodologia *Stratified Cross-Validation*, dividida em 10 partes (Witten e Frank, 2011).

Resultados e Discussão

Na Figura 1 pode-se observar o modelo entidade-relacionamento (MER), sem os campos, criado para melhor compreensão do banco de dados utilizado. Percebe-se pelo MER, a complexidade das tabelas e suas relações. O banco de dados do IFMS, conta com 24 tabelas originais, como mostra o MER. Foi acrescentada uma tabela ao banco, denominada classes, com as seguintes colunas: *campus*, *polo*, *coordenacao*, *curso*, *nome_estudante*, *situacao* (Evadido e Não Evadido), *CPF*, *id*, *curso_id*.

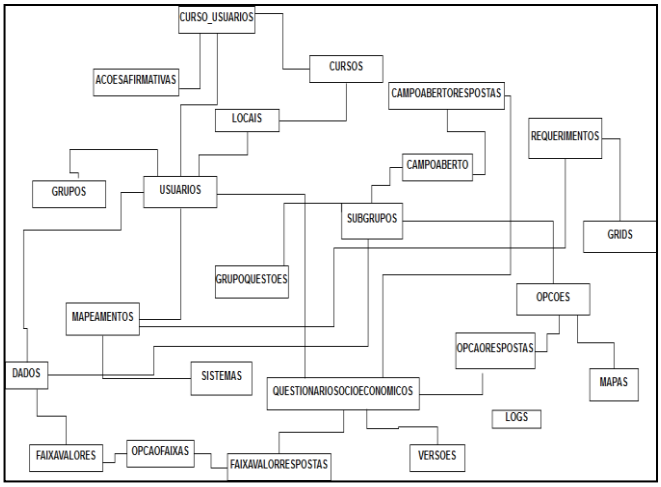

Figura 1. **Modelo Entidade-Relacionamento do Banco de Dados do IFMS**. Fonte: Autoria Própria.

Utilizou-se a ferramenta Weka para gerar e avaliar os classificadores com os quatro algoritmos. Na Tabela 2 pode ser visualizada a relação de acurácia dos algoritmos por campus rodados no Weka. O melhor classificador foi do campus Nova Andradina com o algoritmo J48, enquanto o algoritmo SMO apresenta o pior classificador com o campus Três Lagoas.

	NAIVE BAYES	MULTILAYER PERCEPTRON	SMO	J48
AQUIDAUANA	54,7085%	52,4664%	54,7085%	61,4350%
CAMPO GRANDE	62,5714%	65,7143%	63,1429%	64,8571%
CORUMBÁ	56,6964%	54,9107%	58,4821%	58,9286%
COXIM	73,1935%	62,4709%	70,6294%	71,3287%
IFMS (TODOS OS CAMPI)	64,5755%	56,2735%	63,3734%	62,3591%
NAVIRAÍ	61,0390%	64,9351%	64,9351%	54,5455%
NOVA ANDRADINA	77,0393%	72,8097%	81,8731%	82,1752%
PONTA PORÃ	63,8060%	72,0149%	79,8507%	78,3582%
TRÊS LAGOAS	56,0284%	52,8369%	52,1277%	61,3475%

Tabela 2. **Acurácia dos algoritmos**. Fonte: Autoria Própria.

Foram destacados em azul e vermelho, respectivamente o melhor e o pior resultado de cada Campi, salvo o Campus Naviraí que obteve empate entre os algoritmos *Multilayer Perceptron* e SMO no melhor resultado.

Dos 40 classificadores gerados, o que apresentou melhor desempenho geral foi o algoritmo de Árvore de Decisão J48, tendo 4 melhores acurácias, tendo também a maior média por todos os Campi com 66% de acurácia.

Já o algoritmo *Multilayer Perceptron* obteve os piores resultados, com as 5 piores acurácias, sendo duas (Aquidauana e Corumbá) inferior ao classificador do IFMS geral, que também obteve sua pior acurácia com o *Multilayer Perceptron*. Além disso, a média do algoritmo foi de 61,6%, também o pior resultado das médias dos classificadores analisados.

No Gráfico 1 são apresentadas as acurácias, por classificador e campus, inferiores ao melhor classificador do IFMS geral.

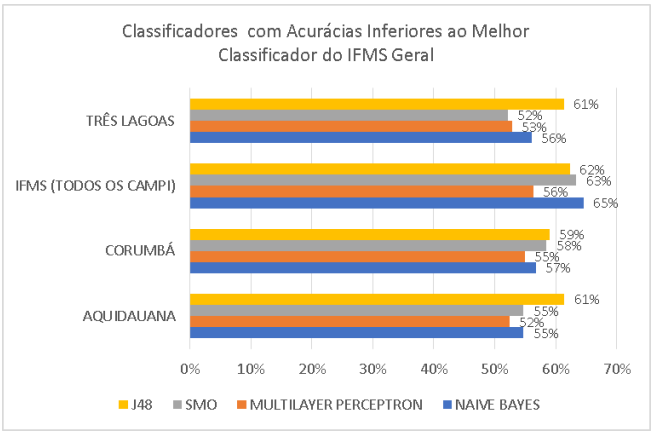

Gráfico 1. **Classificadores com Acurácias Inferiores ao Melhor Classificador do IFMS geral**. Fonte: Autoria Própria.

A melhor acurácia do IFMS geral foi do algoritmo Naive Bayes com 64,5%, enquanto a pior acurácia dos Campi apresentado no Gráfico 1 foi de Três Lagoas do algoritmo SMO com 52,1%. A diferença entre os dois resultados é de 12,4%.

No Gráfico 2 é apresentado as acurácias por campus superiores ao classificador obtido com todos os dados do IFMS.

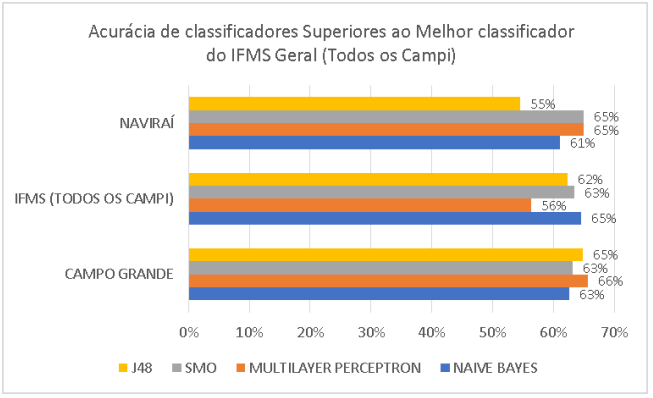

Gráfico 2. **Acurácia de Classificadores Superiores ao Melhor Classificador do IFMS geral (Todos os Campi)**. Fonte: Autoria Própria.

Apesar dos Campi Naviraí e Campo grande obterem uma acurácia superior ao IFMS, percebe-se que os valores não são muito distintos entre si. Enquanto Campo Grande com o algoritmo *Multilayer Perceptron* teve uma acurácia de 65,7%, a melhor acurácia do IFMS geral foi de 64,6%. A diferença entre os dois classificadores é um pouco mais que 1%.

No entanto, no Gráfico 3 poder ser verificado os três Campi cuja as acurácias foram superiores a 70%.

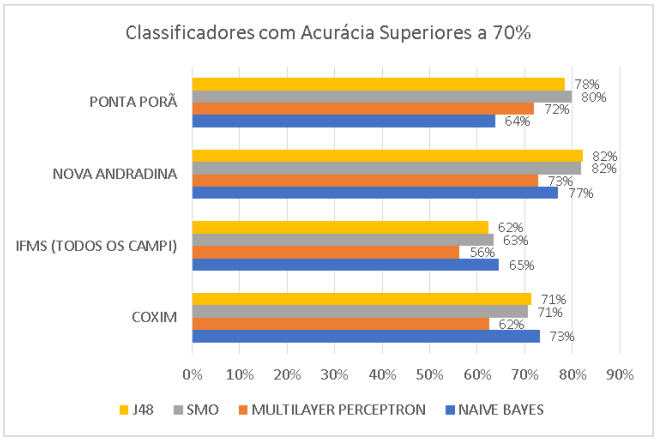

Gráfico 3. **Classificadores com acurácias superiores à 70%.**
Fonte: Autoria própria.

O melhor resultado apresentado no Gráfico 3 foi do Campus Nova Andradina com o algoritmo J48 e 82,2% de acurácia. O pior foi do Campus Coxim com 73,2% de acurácia. A diferença entre os resultados é de 9%. Já a diferença entre o melhor Campus e o IFMS é de 17,6%. A discrepância dos resultados foi superior aos outros propostos. Tanto aos Campus que ficaram abaixo do IFMS, quanto aos que ficaram acima, mas com pouca diferença nas acurácias.

Além das taxas de acertos (apresentadas nos gráficos) e erros, há diversas formas de avaliar o modelo produzido pelos classificadores (AMARAL, 2016). Entre elas, há os verdadeiros positivos. O Gráfico 4 é apresentado os verdadeiros positivos das classes (Evadido e Não Evadido) dos algoritmos com as melhores acurácias para cada Campus.

Os resultados estão em números decimais e vão de 0 a 1. Quanto mais próximo de 0, menos instâncias foram corretamente classificadas, o inverso vale para os valores que se aproximam de 1.

Os verdadeiros positivos mais altos são referentes à classe Não Evadido dos Campi Nova Andradina, Ponta Porã e IFMS, com respectivamente 0,985; 0,979 e 0,944. No entanto os três também apresentam os piores valores para a classe Evadido com 0,018; 0,1 e 0,097. Em quase todos os Campi percebe-se uma melhor classificação dos verdadeiros positivos em relação a Classe Não Evadido. O único Campus que apresenta um resultado inverso é Três Lagoas, onde a classe Evadido tem 0,837 e a Não Evadido tem 0,264.

Uma possível razão para a discrepância dos verdadeiros positivos entre as duas classes pode ser a quantidade de instâncias que os exemplos têm em cada classe. Os Campi Nova Andradina, Ponta Porã e IFMS geral possuem (Evadido e Não Evadido) respectivamente 56 e 275, 110 e 426, e 938 e 1724 instâncias. Tal diferença pode afetar a quantidade de verdadeiros positivos em cada classe, uma vez que quanto maior a quantidade de instâncias mais exemplos os algoritmos podem aprender.

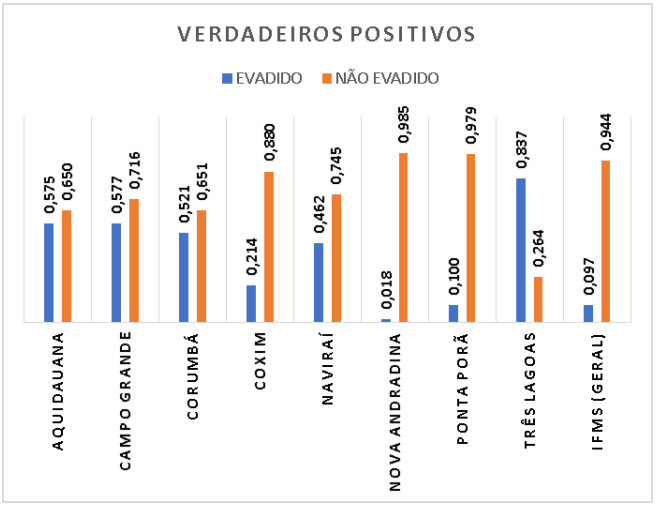

Gráfico 4. **Verdadeiros Positivos**. Fonte: Autoria Própria.

Os modelos gerados foram utilizados para desenvolvimento de Sistema Web para o IFMS (SILVA E SANTOS, 2018 no prelo).

Considerações Finais

O trabalho mostra-se relevante por destacar o uso do Aprendizado de Máquina na identificação de estudantes de cursos superiores que possam evadir de seus cursos, utilizando dados socioeconômicos que são coletados semestralmente no IFMS. Os resultados permitiram identificar a importância de gerar classificadores para cada campus. Os trabalhos futuros poderão melhorar as acurácias dos algoritmos, utilizando outras ferramentas, e também gerar sistemas que proporcionarão que a instituição realize ações específicas para grupos de estudantes com maior possibilidade de desistência de seus cursos.

Agradecimentos

Programa de Iniciação Científica e Tecnológica do IFMS.

Referências

AMARAL, Fernando. **Aprenda Mineração de Dados:** Teoria e Prática. Rio de Janeiro: Alta Books, 2016. 220 p.

CARDOSO, Daniela Freire; LUDOVICO, Nelson. Estudo Longitudinal Sobre As Pesquisas De Evasão No Ensino

Superior: Diretório IBICT. **Refaz**: Revista Fatec Zona Sul, São Paulo, v. 3, n. 4, p.01-18, jun. 2017.

RIGO, Sandro José; BARBOSA, Jorge; CAMBRUZZI, Wagner. Educação em Engenharia e Mineração de Dados Educacionais: Oportunidades para o Tratamento da Evasão. **Revista EAD & Tecnologias Digitais na Educação**, Dourados, v.2, n.3, p.30-40, nov.2014.

Silva, Richard Ribeiro; Santos, Gilson Saturnino. **Desenvolvimento de Aplicação Web Para Apoiar o Controle de Evasão.** Seminário de Iniciação Científica do IFMS, 2018. No prelo.

WITTEN, Ian H.; FRANK, Eibe; MARK, A. Hall. 2011. **Data mining: Practical machine learning tools and techniques**, 2011.