USO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA ANÁLISE DOS FATORES DE SUCESSO DE CURSOS DO ENSINO SUPERIOR NO ENADE

Leandro Guarino de Vasconcelos [1]
Mayara Moura Bento de Castro [2]

RESUMO

No Brasil, a avaliação do desempenho dos alunos de cursos de Ensino Superior é realizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) por meio do Exame Nacional de Desempenho dos Estudantes (Enade). Além das perguntas do exame, os egressos respondem a um questionário sobre o curso e sobre um perfil socioeconômico. O resultado do exame é usado para atribuir a cada curso o chamado Conceito Enade, que é utilizado pelas instituições como forma de atrair candidatos ao vestibular. Neste contexto, a pergunta desta pesquisa foi: o que leva um curso a ter sucesso no Enade? Para isso, foi criado um banco de dados com microdados abertos do Enade de 2018 fornecidos pelo INEP e foram utilizadas análises estatísticas e técnicas de aprendizagem de máquina para encontrar a resposta. Os resultados mostram que as características que influenciam o resultado de um curso no Enade são: a quantidade de alunos participantes do Enade, a média de idade dos alunos, a quantidade de alunos com renda de até 1,5 salários mínimos e de 1,5 a 3 salários mínimos, a quantidade de alunos que escolheu a instituição pelo preço e a quantidade de alunos que escolheu a instituição pela proximidade, pois estes foram os únicos fatores em comum considerados pelos algoritmos ao gerarem suas árvores de decisão.

Palavras-chave: Aprendizagem de máquina, Mineração de dados, Enade.

INTRODUÇÃO

A No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) é responsável por aplicar exames de avaliação para estudantes de ensino médio e superior. O Exame Nacional de Desempenho dos Estudantes (Enade) é o exame responsável por avaliar a qualidade de ensino das instituições de educação superior brasileiras. A partir dos dados obtidos nesse exame, é possível observar e acompanhar o desempenho acadêmico dos estudantes e da instituição de ensino superior, a fim de

[1] Professor Doutor na Faculdade de Tecnologia de Guaratinguetá - FATEC, le.guarino@gmail.com;
[2] Graduando do Curso de Análise e Desenvolvimento de Sistemas da Faculdade de Tecnologia de Guaratinguetá – FATEC, mayara.castro01@fatec.sp.gov.br.

obter os parâmetros necessários para tomadas de decisões de professores, entidades e autoridades educacionais.

O Enade é um exame obrigatório que é aplicado periodicamente, em ciclos de três anos, aos alunos de todos os cursos de graduação. Além dos testes de conhecimento, os alunos devem responder a um questionário sobre perfil socioeconômico e outro sobre a percepção do aluno sobre o teste.

Atualmente, o exame é composto de 180 questões, distribuídas em quatro provas objetivas: Ciências Humanas e suas Tecnologias (História, Geografia, Filosofia e Sociologia); Ciências da Natureza e suas Tecnologias (Química, Física e Biologia); Linguagens, Códigos e suas Tecnologias (Língua Portuguesa, Literatura, Língua Estrangeira – Inglês ou Espanhol, Artes, Educação Física e Tecnologias da Informação e Comunicação); Matemática e suas Tecnologias (Matemática). Além disso, os alunos devem escrever uma redação.

Os resultados obtidos no exame são divulgados na forma de relatórios de Curso, de IES e Síntese de Área, com estatísticas geradas a partir dos dados. Esses dados informam o desempenho dos estudantes, as suas percepções sobre a prova e as estatísticas das questões da prova e os resultados da Análise do Questionário do Estudante.

O Inep realiza análises que geralmente se resumem a estatísticas descritivas dos dados e que visam resumir as informações obtidas. Análises que propõem extrair informações mais profundas a partir de um grande volume de dados não são divulgadas pelo Inep. Portanto, com o objetivo de obter informações relevantes sobre o que leva os cursos de ensino superior ao sucesso no Enade, esta pesquisa utilizou-se de técnicas de aprendizagem de máquina para detectar padrões entre as características dessas instituições e de seus respectivos cursos, utilizando os dados abertos disponibilizados pelo Inep e pelo MEC.

1.1 Inteligência Artificial

Inteligência Artificial é o ramo da Ciência da Computação que busca desenvolver sistemas de computadores inteligentes que simulem e que exibam características da inteligência natural do homem. São exemplos dessas características: raciocínio lógico, aprendizagem, resolução de problemas e compreensão da linguagem. De acordo com Santos (2005), um processo de aprendizagem inclui novas formas de conhecimento: o desenvolvimento motor e a habilidade cognitiva, a organização do

novo conhecimento e descobertas de novos fatores e teorias através da observação e experimentação.

1.2 Aprendizagem de Máquina

Aprendizagem de Máquina (AM) ou Aprendizagem Computacional é uma sub-área da Inteligência Artificial que estuda métodos computacionais para adquirir novos conhecimentos e novas habilidades. Existem vários métodos de aprendizagem de máquina, como por exemplo, a aprendizagem por hábito, por instrução, por dedução, por analogia e por indução.

As técnicas de AM podem ser divididas em duas classes em geral: aprendizagem supervisionada e aprendizagem não supervisionada. A aprendizagem supervisionada consiste no algoritmo de aprendizagem (indutor) receber um conjunto de exemplos, e cada exemplo ser formado por um conjunto de atributos de entrada e atributos de saída, ou seja, o indutor recebe um conjunto de exemplo para treinamento para os quais os rótulos da classe já são conhecidos. Já na aprendizagem não supervisionada, o indutor recebe um conjunto de exemplos formados por conjuntos de atributos de entrada somente, ou seja, o indutor analisa os exemplos fornecidos para tentar agrupá-los de alguma maneira. Essas técnicas são utilizadas para buscar e detectar padrões em uma grande quantidade de dados que auxiliem no entendimento dos mesmos.

1.3 Aprendizagem de Máquina aplicada à educação

Existem na literatura alguns artigos referentes à aprendizagem de máquina aplicada na área de educação. Amorim (2008) aplicou técnicas de AM para fazer uma previsão de evasão acadêmica, concentrando-se na modelagem dos principais aspectos que podem levar um aluno a trancar ou abandonar seu curso. Nesse artigo, foram implementadas três fases principais para a criação de um sistema de previsão, sendo elas: seleção de atributos, levantamento dos dados e escolha dos classificadores, testando a acurácia dos mesmos e, por último, mostrou as estatísticas referentes à evasão de cada curso.

Brito et al. (2014) propuseram a utilização de técnicas de Mineração de Dados para tentar correlacionar as notas de ingresso dos alunos no curso superior aos seus respectivos desempenhos no primeiro período do curso de Ciência da Computação da UFPB e, dessa forma, obter informações para a realização de ações contra a evasão acadêmica.

A fim de analisar a infraestrutura das escolas, os perfis dos estudantes e os perfis das instituições de ensino superior do estado de Pernambuco, Carvalho et al. (2017) aplicaram algoritmos de aprendizagem de máquina para o processo de mineração de dados abertos do Inep referentes aos Censos da Educação Básica e Superior.

Moriconi et al. (2014) buscaram identificar fatores associados ao desempenho de novos engenheiros no Enade. Para isso, fez uso dos dados do Enade 2011 e de modelagens hierárquico-lineares. O artigo apresenta resultados que reforçam a percepção geral de que cursos de instituições públicas se destacam na formação de engenheiros, além de que os alunos que cursaram o ensino superior na idade correta, que fizeram ensino médio em escolas públicas e alunos do sexo masculino e que se declararam brancos, obtiveram o melhor desempenho na prova.

Ferreira (2015) teve como objetivo identificar as variáveis significativas na explicação do resultado do Enade. Para tal, selecionou uma amostra de 77% do total dos alunos de Ciências Contábeis que realizaram a prova em 2012. Por meio do uso de estatística descritiva, encontrou que as variáveis significativas foram: gênero, estado civil, etnia, renda, bolsa de estudo, forma de ingresso, escolaridade da mãe, escola ensino médio (pública ou privada), tipo de ensino médio (tradicional ou profissionalizante), quantidade de livros, horas estudadas, participação iniciação científica, participação monitoria, participação atividades de extensão, categoria administrativa da IES, região, número de concluintes participantes do Enade, nota de ingressantes (Enem), percentual de mestres, percentual de infraestrutura, percentual organização didático-pedagógica.

O trabalho de Silva et al. (2015) estudaram os fatores que impactam o desempenho dos estudantes de Administração na nota do Enade. Por meio de análise fatorial e regressão múltipla, encontraram pouca explicação para o desempenho dos alunos nas provas, e identificaram que os resultados encontrados estão direcionados para a maturidade do aluno e suas bases educacionais.

Cretton & Gomes (2016) tiveram como objetivo extrair conhecimento do curso de medicina através da análise dos dados de 2013 do Enade. Utilizaram o método KDD (Knowledge Discovery in Databases), técnicas de mineração de dados juntamente com o *software Weka*, e então observaram a influência da categoria e dos tipos das IES no resultado do exame.

Crepalde et al. (2016) analisaram a relação das desigualdades educacionais com o desempenho dos alunos no Enade 2014, através de modelos lineares e modelos

hierárquicos. Encontraram que o efeito das escolas sobre o desempenho dos alunos supera as diferenças de desempenho por sexo, raça e renda familiar, além do comportamento distintivo em cada curso quando se levam em conta as desigualdades.

O objetivo no trabalho de Rocha et al. (2018) foi verificar a associação entre o desempenho dos estudantes de Nutrição no Enade e fatores socioeconômicos, trajetória acadêmica e perfil da instituição. Realizaram análise descritiva, regressão linear simples e múltipla, e encontraram que a categoria administrativa da IES foi o principal fator associado ao desempenho no Enade.

METODOLOGIA

A metodologia A metodologia desta pesquisa consiste na pesquisa bibliográfica referentes aos conceitos utilizados na pesquisa, e em uma pesquisa quantitativa a partir dos dados públicos disponibilizados pelo Inep e pelo MEC.

Para alcançar o objetivo, foi utilizado o processo KDD (Knowledge Discovery in Databases), que consiste em uma sequência de etapas que devem ser executadas sequencialmente. Essas etapas são: seleção dos dados, pré-processamento e limpeza, transformação dos dados, mineração dos dados, interpretação e avaliação, conforme ilustra a Figura 1.

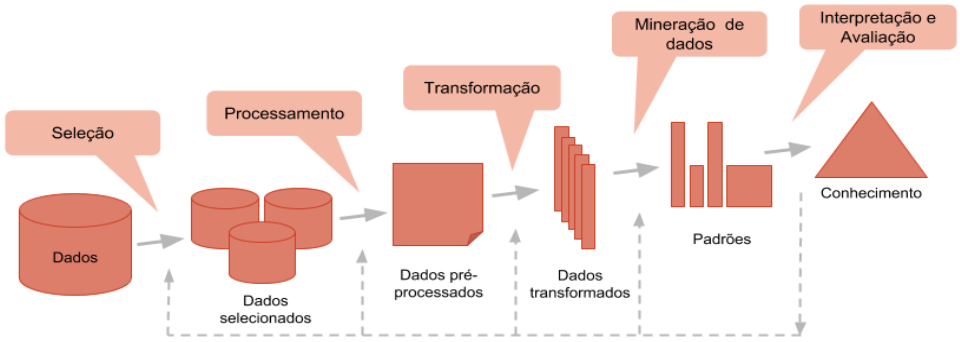

Figura 1: Processo KDD.

1.1 Primeira etapa: Criação do Banco de Dados

O primeiro passo foi acessar o site do Inep, onde são disponibilizados os dados do Censo da Educação Superior, como informações sobre as IES, os cursos oferecidos, locais, alunos e os dados das provas do Enade realizadas nos anos anteriores. Neste trabalho, foram utilizados os dados do Enade do ano de 2018, que avaliou cursos de 27

áreas, dentre elas 4 áreas de cursos oferecidos pela FATEC Guaratinguetá: Tecnologia em Gestão Empresarial, Tecnologia em Gestão Comercial, Tecnologia em Gestão Financeira e Tecnologia em Logística.

Os microdados do Enade 2018 estão publicados em arquivos no formato CSV (*Comma Separated Values*, em inglês). Para facilitar a manipulação desses dados, foi criado um banco de dados utilizando o Sistema Gerenciador de Banco de Dados MySQL com as mesmas tabelas existentes nos arquivos CSV. Para inserir os dados dos arquivos CSV no banco de dados criado, foi escrito um script em JavaScript para cada arquivo. O script é executado utilizando o ambiente de execução Node.js e conecta-se ao banco de dados MySQL para inserir os dados dos arquivos CSV.

O desafio nessa etapa foi ler arquivos com centenas de milhares de linhas e inseri-las no banco de dados no menor tempo possível. Para alcançar um bom resultado, foram estudados recursos de leitura assíncrona de arquivos.

Com os dados já inseridos no banco de dados, foi necessário converter os tipos de colunas cujo conteúdo é numérico, pois os dados foram inseridos no formato texto. Após a conversão, o próximo passo foi verificar dados que podiam ser ignorados devido a valores inexistentes ou irrelevantes. Portanto, os campos que apresentavam valor "NA" foram substituídos por valor NULL.

Finalizando essa etapa, foram criados índices nas tabelas do banco de dados, que permitem acelerar a execução de consultas em tabelas com grande conjunto de dados. Isso foi necessário pois a tabela de alunos, por exemplo, possui mais de 12 milhões de linhas (registros).

1.2 Segunda etapa: Consultas no Banco de Dados

Para dar início ao processo de descoberta de conhecimento em bases de dados, foi necessário pensar quais seriam os atributos das tabelas que permitiriam encontrar boas relações entre os dados, e que poderiam retornar resultados que contribuíssem para detectar padrões e comportamentos.

Essa etapa é importante que se possa explorar os dados e conhecê-los, pois esse conhecimento permitirá interpretar os padrões que serão detectados ao final do processo de mineração de dados.

Em seguida, para colocar em prática as ideias do passo anterior, foram criados comandos para realizar as seguintes consultas:

- Contar a quantidade de alunos participantes do Enade 2018 e a quantidade de participantes com resultados válidos;
- Calcular a média da Nota Geral para cada curso de cada IES, para os participantes presentes e com resultados válidos;
- Calcular a média nacional da Nota Geral para cada curso, para os participantes presentes e com resultados válidos;
- Calcular a quantidade de alunos, a média da Nota Geral e média da Nota de Conhecimentos Específicos para os cursos de Processos Gerenciai;
- Calcular a quantidade de alunos que fizeram o Enade, a quantidade de alunos que obtiveram resultados abaixo da média geral e a quantidade de alunos que obtiveram resultados acima da média geral dos cursos da área de Processos Gerenciais;
- Calcular a média geral dos alunos para cada curso da IES, a média da Nota Geral dos alunos para cada curso da IES, média da Nota de Conhecimento Específico dos alunos para cada curso da IES – área de Processos Gerenciais;
- Calcular a quantidade de IES que obteve resultados acima da média geral nacional e a quantidade de IES que obteve resultados abaixo da média geral nacional.

2.3 Terceira etapa: Criação da tabela para mineração dos dados

A tabela de dados do Enade do ano de 2018 não possui uma coluna com os valores do Conceito Enade para cada curso, oferecendo somente colunas para as notas que compõem o resultado final da prova de cada aluno. Dessa forma, foi necessário conhecer o método utilizado pelo Inep para o cálculo do Conceito Enade, que varia de 1 a 5, e aplicá-lo utilizando a linguagem SQL, guardando os resultados em uma coluna da tabela Cursos, no banco de dados. Os cálculos realizados pelo Inep estão disponíveis em uma nota técnica publicada pelo Instituto[3]. Para facilitar a aplicação dos cálculos, foi criada uma coluna "co_grupo" na tabela Cursos, que foi preenchido com os valores da coluna "co_grupo" da tabela com os dados do Enade de 2018, para identificar a categoria de curso a que o grupo pertence. Também foram criados os campos para o afastamento padronizado, notas padronizadas e máximos e mínimos do afastamento padronizado – que foram preenchidos com os resultados dos cálculos indicados pelo

3 http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2017/Nota_Tecnica_CGCQES_n12_2017_Calculo_da_nota_final_do_Enade.pdf

Inep. Por fim, com os campos que contêm os valores necessários para o cálculo final da nota já gerados, o campo “conceito enade 2018” foi criado para armazenar o conceito Enade de cada curso. É importante destacar que o conceito Enade de um curso é relativo às notas de Formação Geral e Conhecimentos Específicos dos outros cursos da mesma área, gerando uma nota padronizada com valores de 1 a 5.

Continuando com o pré-processamento para explorar os dados dos cursos de todo o país, foi criada uma tabela com os atributos considerados mais relevantes para se obter uma correlação com o desempenho no conceito Enade e para a posterior mineração desses dados.

A penúltima etapa do pré-processamento foi remover os valores fora dos limites de controle. Portanto, para cada atributo numérico, os valores superiores à somatória da média e do desvio-padrão foram preenchidos como nulos, para não influenciarem na análise dos dados, pois são valores conhecidos como *outliers*.

Para finalizar o pré-processamento, foi incluído nessa tabela o campo “sucesso_enade”, que possui um valor binário: Sucesso e fracasso. Os cursos que obtiveram conceito Enade >= 3 foram considerados como sucesso, do contrário, foram considerados como fracasso.

2.4 Quarta etapa: Aplicação das técnicas de Aprendizagem de Máquina

Na etapa anterior, foi criada a tabela que foi usada posteriormente para a mineração de dados, que contém os atributos considerados relevantes para a obtenção de sucesso ou fracasso no Enade. Nesta última etapa, o objetivo foi usar algoritmos de Aprendizagem de Máquina (AM) para gerar modelos de predição e classificação para os dados do Enade e, assim, prever se determinado curso obter as regras que levam um curso ao sucesso ou fracasso na prova.

A tabela preenchida anteriormente foi importada em formato CSV no *software* Weka (*Waikato Environment for Knowledge Analysis*), versão 3.8.4. O processo de mineração de dados desse software consiste na aplicação de algoritmos de AM para a regressão, classificação, agrupamento ou seleção de atributos dos dados, dentre outros (RUIZ et al., 2018).

O método de AM escolhido foi o método supervisionado chamado árvore de decisão. No campo da ciência da computação, árvores são estruturas de dados formadas por um conjunto de elementos que armazenam informações chamadas nós (MEDIUM, 2017). Os nós representam regiões onde são realizados testes lógicos para a separação

dos dados - o primeiro nó é chamado de nó raiz e é o nó principal da árvore de decisão (SATO et al., 2013).

Para mineração dos dados, foram executados todos os algoritmos disponíveis no *software* Weka referentes ao método de árvore de decisão, pois o objetivo era encontrar qual algoritmo aprende melhor os padrões de acordo com os dados pré-processados.

Na aprendizagem supervisionada, os dados são fornecidos com os valores de entrada e saída para que o algoritmo gere um classificador, ou seja, um classificador é treinado para aprender padrões sobre os dados de entrada. Dessa forma, 66% dos dados da tabela para mineração foram usados para o treinamento dos algoritmos, enquanto os 34% restantes foram usados para validar o modelo de treinamento gerado pelo algoritmo.

Para comparação dos resultados, foram considerados os seguintes indicadores:

- Porcentagem de instâncias classificadas corretamente;
- Quantidade de instâncias classificadas como “Sucesso”;
- Quantidade de instâncias classificadas como “Fracasso”.

Após comparar esses indicadores, os algoritmos que tiveram melhor desempenho foram o *Random Forest*, LMT (*Logistic Model Tree*) e o REPTree.

RESULTADOS E DISCUSSÃO

Nos A partir dos resultados das consultas realizadas no banco de dados, foram criadas planilhas no software Excel da Microsoft, e assim foi possível realizar uma análise mais minuciosa desses resultados, referentes ao Enade 2018.

Na Figura 1, é possível notar que o número total de cursos que ficaram acima da média nacional foi de 1988 cursos, e o número total de cursos que obtiveram resultados abaixo da média nacional foi de 4825 cursos, ou seja, mais de 70% dos cursos do Brasil têm nota inferior à média nacional do respectivo curso.

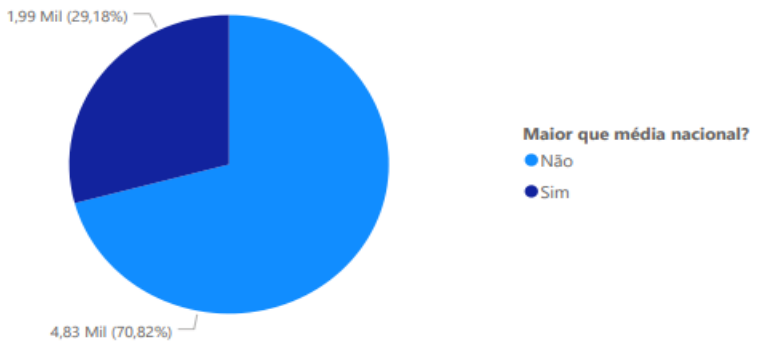

Figura 2: Gráfico da porcentagem de cursos com média acima e abaixo da média nacional dos cursos avaliados no Enade 2018.

A Figura 2 mostra a média geral nacional de cada curso. Observa-se que a maior média nacional é do curso de Administração, enquanto a menor média nacional é do curso de Comunicação Social - Publicidade e Propaganda.

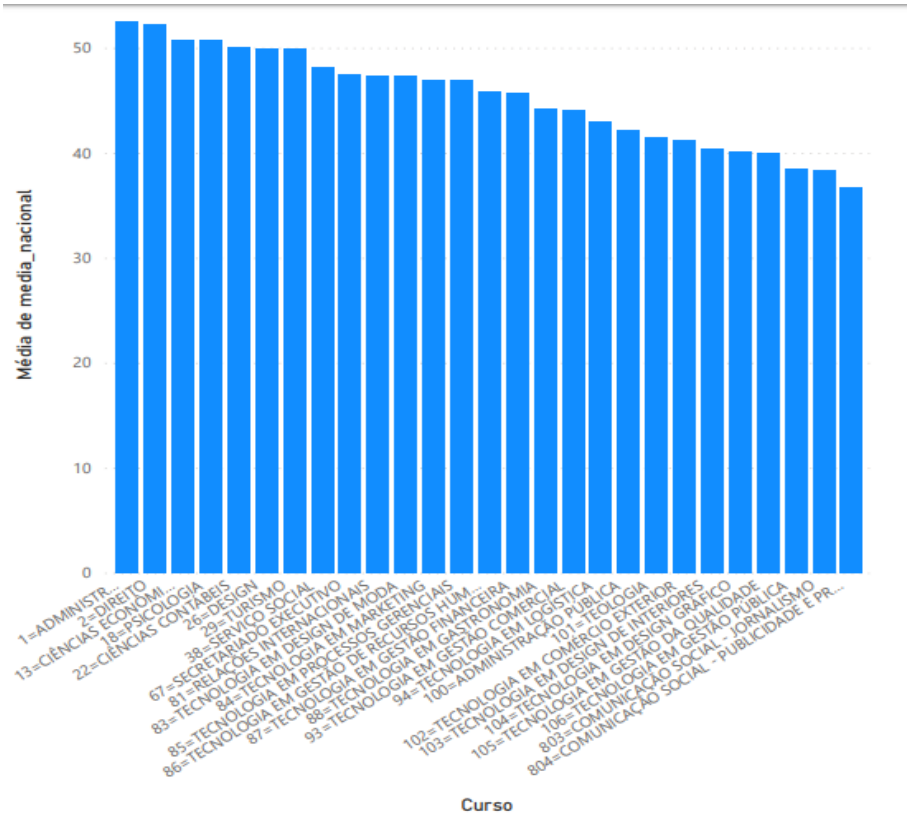

Figura 3: média nacional de cada curso avaliado no Enade 2018.

3.1 Aplicação das técnicas de aprendizagem de máquina

Para a análise de todos os cursos de ensino superior das IES brasileiras participantes do Enade 2018, a partir da tabela criada com os atributos mais relevantes dos cursos, foi utilizado o método supervisionado de aprendizagem de máquina, mais especificamente a técnica de classificação.

Como os algoritmos dessa técnica se baseiam na correlação entre os atributos dos dados de entrada, foi realizado um estudo prévio da correlação entre os atributos de entrada e a nota do curso no Enade 2018. A nota utilizada foi a nota antes da normalização aplicada para gerar o conceito do curso no Enade (de 1 a 5).

Para a análise de correlação, foi calculado o coeficiente de correlação de Pearson entre cada atributo numérico e a nota do curso no Enade 2018. A Tabela 5 apresenta os 10 atributos com maior coeficiente de correlação.

Tabela 1 - Atributos selecionados pelos algoritmos para definição dos padrões

Variável	**Correlação**
Quantidade de alunos com mãe com ensino superior	0,3821993592
Quantidade de alunos com pai com ensino superior	0,3663149244
Quantidade de alunos cuja renda familiar está entre 10 e 30 salários mínimos	0,2917025882
Quantidade de alunos com alguém da família com ensino superior	0,2645412438
Quantidade de alunos sem renda sustentados por programa governamental ou pela família	0,2337834106
Quantidade de alunos que escolheram a IES por proximidade	-0,2318337873
Despesa com docente por aluno	0,2233921497
Quantidade de alunos cuja renda familiar está entre 5 e 10 salários mínimos	0,2195358166
Quantidade de alunos cuja renda familiar é de 1,5 a 3 salários mínimos	-0,2156553242
Média de idade dos alunos	-0,2074250009

Observa-se na Tabela 1 que os atributos que possuem maior correlação com a nota do Enade são atributos relacionados à estrutura familiar e à renda. No entanto, observa-se que a correlação entre os atributos individuais e a nota do curso no Enade 2018 é fraca (menor que 0,5), ou seja, não é possível afirmar que há uma correlação entre um atributo e a nota do curso. Isso mostra a complexidade do problema investigado nesta pesquisa. Como os dados são aleatórios em relação à nota do curso no Enade 2018, faz-se necessária a exploração das técnicas de aprendizagem de máquina para construir um modelo de classificação a partir dos dados observados.

Para isso, com o auxílio do software *Weka*, foi possível selecionar os algoritmos de AM para realizar a mineração dos dados dos cursos de todo o país. Os algoritmos selecionados foram:

- *Random Forest*: *random* significa aleatório, e denota o comportamento do algoritmo ao selecionar subconjuntos de *features* e montar mini árvores de decisão. *Forest* significa floresta, já que são geradas várias árvores de decisão.
- LMT: combina regressão logística e aprendizagem de árvore de decisão. De acordo com Ruiz et al. (2018), a característica marcante da árvore de decisão gerada pelo algoritmo LMT é a sua estrutura com funções de regressão logística nas folhas, e em seus nós terminais, ao invés de classes únicas, possui vetores de variáveis (xi) e de coeficientes (βj) de todas as classes. Através desses valores, é possível calcular a função LMT.
- REPTree: constrói árvores de decisão para classificação ou regressão com base no ganho de informação/variância e poda esta árvore usando uma poda guiada por erro (WITTEN et al., 2011).

4.1.1 Random Forest

O algoritmo *Random Forest* classificou 83.741% das instâncias corretamente. Na Figura 3, observa-se a matriz de confusão gerada pelo algoritmo. Pode-se observar que o número de instâncias classificadas corretamente foi consideravelmente maior para o caso “a” (= sucesso). Isso significa que esse algoritmo aprendeu bem o padrão de “sucesso”, mas classificou corretamente apenas 36 instâncias quanto ao “fracasso”.

```
=== Confusion Matrix ===

   a    b   <-- classified as
2292   15 |    a = SUCESSO
 437   36 |    b = FRACASSO
```

Figura 4: Matriz de Confusão referente ao algoritmo Random Forest

4.1.2 LMT

O algoritmo LMT classificou 83.2014% de instâncias corretamente, sendo essa uma porcentagem um pouco maior que a do algoritmo *Random Forest*. A maior diferença entre os dois algoritmos está em suas matrizes de confusão. Como é possível ver na Figura 4, a matriz de confusão gerada pelo LMT obteve mais acertos para o caso "b" (= fracasso) quando comparado com o algoritmo *Random Forest*, porém, obteve um resultado um pouco pior para o caso "a" (= sucesso).

```
=== Confusion Matrix ===

   a    b   <-- classified as
2263   44 |    a = SUCESSO
 423   50 |    b = FRACASSO
```

Figura 5: Matriz de Confusão referente ao algoritmo LMT.

A árvore de decisão gerada pelo LMT indica os atributos considerados mais determinantes pelo algoritmo, e são eles: a média de idade dos alunos participantes do Enade 2018 e a renda dos alunos participantes. Essa árvore pode ser vista na Figura 5.

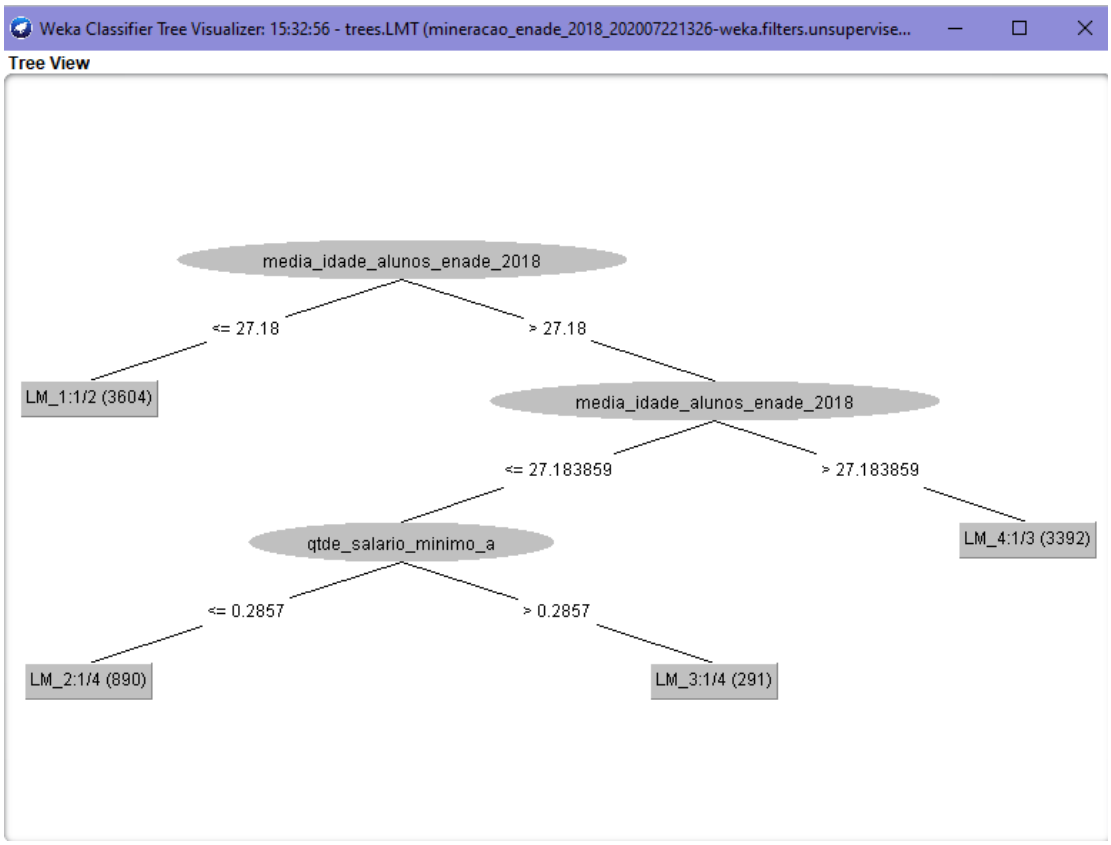

Figura 6 - Árvore de decisão gerada pelo algoritmo LMT.

4.1.3 REPTree

O algoritmo REPTree conseguiu classificar corretamente 81.3309% das instâncias. Foi o algoritmo que atingiu a menor porcentagem de acertos dentre os três algoritmos testados neste trabalho. Na sua matriz de confusão, percebe-se que aconteceu algo semelhante ao resultado do algoritmo LMT. O número de acertos para o caso "fracasso" foi igual, e melhor que o resultado obtido pelo algoritmo *Random Forest*, porém, para o caso "sucesso", o algoritmo obteve o pior desempenho entre os três.

```
=== Confusion Matrix ===

    a    b   <-- classified as
 2211   96 |    a = SUCESSO
  423   50 |    b = FRACASSO
```

Figura 7 - Matriz de confusão referente ao algoritmo REPTree.

4.1.4 Atributos Relevantes

Devido à complexidade das árvores geradas pelos algoritmos *Random Forest* e REPTree, não é possível apresentar uma figura que mostre a árvore completa. Portanto, para visualizar quais atributos de entrada foram selecionados por esses algoritmos para determinar o sucesso ou o fracasso no Enade 2018, a Tabela 2 mostra quais foram os atributos que cada algoritmo levou em conta para realizar a classificação de um curso como sucesso ou fracasso. O atributo que apareceu em algum momento nas árvores de decisão geradas pelos algoritmos aparece marcado com um "X".

Tabela 2 - Atributos selecionados pelos algoritmos para definição dos padrões

Atributo	***Random Forest***	**LMT**	**REPTree**
Categoria administrativa	X		X
Organização acadêmica	X		X
Modalidade de ensino	X		X
Demanda de ingresso total	X		X
Receita própria da IES por aluno	X		X
Despesa da IES com investimento por aluno	X		X
Despesa com docente por aluno	X		X
Quantidade de alunos participantes Enade 2018	X	X	X
Média de idade dos alunos	X	X	X
Média do tempo até entre Ensino Médio e Início da graduação	X		X
Quantidade de alunos com pais com ensino superior	X		

Quantidade de alunos com mães com ensino superior	X		X
Quantidade de alunos com renda até 1,5 salários mínimos	X	X	X
Quantidade de alunos com renda de 1,5 até 3 salários mínimos	X	X	X
Quantidade de alunos com renda de 3 até 4,5 salários mínimos	X		X
Quantidade de alunos com renda de 4,5 até 6 salários mínimos	X		X
Quantidade de alunos com renda de 6 até 10 salários mínimos	X		
Quantidade de alunos com renda de 10 até 30 salários mínimos	X		X
Quantidade de alunos com renda acima de 30 salários mínimos	X		
Quantidade de alunos sem renda	X		X
Quantidade de alunos que sustentam a família	X		X
Quantidade de alunos que leram livro além da bibliografia do curso	X		X
Quantidade de alunos com alguém da família com ensino superior	X		X
Quantidade de alunos com dedicação parcial ao curso	X		X

Quantidade de alunos que apenas assistiram a aulas	X		X
Quantidade de alunos com dedicação total ao curso	X		
Quantidade de alunos que escolheram IES pela gratuidade	X		
Quantidade de alunos que escolheram IES pelo preço	X	X	X
Quantidade de alunos que escolheram IES pela proximidade	X	X	X

O *Random Forest* foi o único algoritmo a considerar todos os atributos como relevantes. Como seu algoritmo gera mais de uma árvore de decisão, e neste trabalho ele foi configurado no Weka para gerar 50 árvores, as possibilidades encontradas por ele foram maiores que a dos outros dois algoritmos. O algoritmo LMT considerou apenas seis atributos como relevantes para suas predições, sendo que, em sua árvore de decisão mostrada na Figura 5, ele considerou apenas a média de idade e a quantidade de alunos com renda de até 1,5 salários mínimos como sendo as únicas relevantes. O algoritmo REPTree levou em conta a maioria dos atributos, deixando apenas cinco deles fora de sua árvore de decisão. Portanto, pode-se concluir que as regressões logísticas realizadas pelo algoritmo LMT se mostram bastante eficientes para a predição desde tipo de problema, já que com um número de atributos consideravelmente menor o algoritmo mostrou uma eficiência muito próxima ao *Random Forest*, que fez uso de todos os atributos para gerar suas árvores de decisão e atingiu a maior porcentagem de instâncias classificadas corretamente dentre os três.

CONSIDERAÇÕES FINAIS

O objetivo inicial deste trabalho foi determinar características dos cursos e das IES brasileiras que pudessem ser relevantes e determinísticas para o sucesso ou

fracasso de seus cursos no Enade. Após todo o tratamento dos dados abertos disponibilizados pelo Inep sobre as instituições, alunos, cursos, locais de ofertas e do Enade 2018, foi possível encontrar algumas estatísticas dos cursos oferecidos pelas IES, e concluir que mais da metade dos cursos avaliados em 2018 possuem notas abaixo da média nacional.

Para tentar encontrar os motivos que levam a tal estatística, foram executados algoritmos de aprendizagem de máquina supervisionados no *software* Weka, com os atributos selecionados que pareciam ser relevantes para um determinado curso obter sucesso ou fracasso na prova. Com esses atributos, os algoritmos puderam aprender a classificar os cursos como sucesso ou fracasso, com aproximadamente 83% de acurácia, e gerar as árvores de decisão que mostram os atributos considerados como relevantes para as classificações. Essas árvores de decisão podem ser utilizadas para classificar outros cursos em próximas edições do Enade, a fim de descobrir se um curso terá sucesso ou fracasso na prova.

Por meio da análise de correlação linear, foi possível observar que os atributos mais relacionados à nota do Enade 2018 referem-se à estrutura e à renda familiar. No entanto, não há uma correlação entre um atributo individualmente, tendo em vista que os coeficientes de correlação são menores que 0,5.

Essa aleatoriedade dos dados levou os algoritmos *Random Forest* e REPTree a consideraram todos ou quase todos os atributos de entrada na criação das árvores de decisão. Já o algoritmo LMT mostrou que, com apenas alguns dos atributos selecionados, é possível uma obter uma classificação muito semelhante à classificação dos outros algoritmos.

A partir dessas análises, conclui-se que a quantidade de alunos participantes do Enade 2018, a média de idade, a quantidade de alunos com renda de até 1,5 salários mínimos e de 1,5 a 3 salários mínimos, a quantidade de alunos que escolheu IES pelo preço e a quantidade de alunos que escolheu a IES pela proximidade, foram os únicos a serem considerados pelos três algoritmos, portanto são bastante relevantes para a correlação com o sucesso ou fracasso dos cursos superiores avaliados pelo Enade em 2018. Analisando esses atributos em comparação com os coeficientes de correlação linear, conclui-se que a média de idade dos alunos do curso, a quantidade de alunos com baixa renda e a quantidade de alunos cuja escolha da IES foi por proximidade ou por preço possuem correlação negativa com a nota na prova do Enade, enquanto a

quantidade de alunos participantes na prova apresentou coeficiente igual a 0,00765, que representa uma correlação nula com a nota da prova.

REFERÊNCIAS

AMORIM, J. V. Maurício. Técnicas de Aprendizagem de Máquina Aplicada na Previsão de Evasão Acadêmica. 2008.

BRITO, Daniel Miranda et al. Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. 2014.

CAMPOS, Raphael. Árvores de Decisão. Medium, 28 de nov. de 2017. Disponível em <https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69>

CARVALHO, Jonathan H. A. et al. Descoberta de conhecimento com Aprendizado de Máquina Supervisionado em Dados Aberto dos Censos da Educação Básica e Superior. 2017.

CREPALDE et al. Desempenho universitário no Brasil: estudo sobre desigualdade educacional com dados do Enade 2014. Revista Brasileira de Sociologia | Vol. 04, No. 07 | Jan. Jun, 2016.

CRETTON & GOMES. Aplicação de técnicas de mineração de dados na base de dados do ENADE com enfoque nos cursos de Medicina. Acta Biomedica Brasiliensia / Volume 7/ no 1/ Julho de 2016.

FERREIRA et al. Determinantes do desempenho discente no ENADE em cursos de Ciências Contábeis. 2015.

MORICONI et al. Fatores associados ao desempenho dos concluintes de engenharia no ENADE 2011. Est. Aval. Educ., São Paulo, v. 25, n. 57, p. 248-278, jan./abr. 2014.

ROCHA et al. Fatores associados ao desempenho acadêmico de estudantes de Nutrição no Enade. Rev. Bras. Estud. Pedagog. vol.99 no.251 Brasília Jan./Apr. 2018.

RUIZ, P. R. S. Classificação da cobertura do solo urbano usando árvores de decisão a partir de cenas WorldView-2 e WorldView-3 para diferentes níveis de legenda. 2017. 203.p. IBI: <8JMKD3MGP3W34P/3NJ9GU8>. (sid.inpe.br/mtc-m21b/2017/03.23.16.13-TDI). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2017. Disponível em: <http://urlib.net/rep/8JMKD3MGP3W34P/3NJ9GU8>.

SANTOS, Cícero Nogueira dos. Aprendizagem de máquina na identificação de sintagmas nominais: o caso do português brasileiro. Rio de Janeiro, 2005.

SATO, L. Y.; SHIMABUKURO, Y. E.; KUPLICH, T. M.; GOMES, V. C. F. Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação do uso e cobertura da terra. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 16. (SBSR), 2013, Foz do Iguaçu. Anais... São José dos Campos: INPE, 2013. p. 2353-2360. DVD, Internet. ISBN 978-85-17-00066-9 (Internet), 978-85-17-00065-2 (DVD). IBI: <3ERPFQRTRW34M/3E7GFLK>. Disponível em: <http://urlib.net/rep/3ERPFQRTRW34M/3E7GFLK>.

SILVA et al. Fatores determinantes para o desempenho dos alunos de administração no ENADE. XV COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA – CIGU. 2015.

WITTEN et al. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Burlington, MA, 2011, 3 ed.