Apoena Revista Eletrônica

ANÁLISE DE FATOR DE EVASÃO: O impacto do tempo na tomada decisão de alunos

Filipe de Lima Venturi

Celso Barreto da Silva

Fabio Fonseca Barbosa Gomes

José Vicente Cardoso Santos

Resumo: O artigo conflitua diferentes modelos de aprendizagem de máquina na análise de dados dos alunos que demonstraram interesse na evasão do seu curso superior em decorrência do impacto do tempo na tomada das suas decisões sobre seu futuro acadêmico. A priori, será apresentado uma fundamentação teórica sobre os tipos de aprendizado de máquina e, em seguida, os diferentes modelos utilizados na pesquisa, o que inclui a *decision tree, random forest* e *neural network*. Por fim, serão expostos os resultados para explicar a escolha do método utilizado para apurar com maior eficácia os fatores externos e internos que contribuem para evasão à medida que o aluno avança seus semestres. Este trabalho tem como objetivo realizar uma análise dos dados educacionais disponibilizados pela mineração de dados no mestrado em Assessoria de Administração do ISCAP em Porto - PT e o dataset interno da Universidade Jorge Amado (UNIJORGE), a fim de comprovar a assertividade da variável tempo de curso na previsão de evasão escolar no ensino superior.

Palavras-chave: Impacto do Tempo. Análise de Dados. Aprendizagem de Máquina.

Abstract: The article clashes different models of machine learning in the analysis of data from students who showed interest in dropping out of their graduation course due to the impact of time in making their decisions about their academic future. Firstly, a theoretical foundation on the types of machine learning will be presented and then the different models used in the research, which include the decision tree, random forest and neural network. Finally, the results will be exposed to explain the choice of method used to more effectively determine the external and internal factors that contribute to dropout as the student progresses through their semesters.This work aims to carry out an analysis of the educational data made available by data mining in the master's degree in Administration Advisory at ISCAP in Porto - PT and the internal dataset of the Universidade Jorge Amado (UNIJORGE), in order to prove the assertiveness of the course time variable in the forecast dropout rates in graduation.

Keywords: Impact of Time. Data Analysis. Machine Learning.

1 INTRODUÇÃO

A evasão escolar trata-se de um problema abordado mundialmente por diversos pesquisadores, uma vez que representa um obstáculo na classificação e reputação das universidades.

Tal fato se acentua mais ainda quando se trata de universidades públicas, posto que representa um prejuízo sobre os investimentos financeiros destinados ao ensino superior.

Conforme Silva, Cabral e Pacheco (2020), a evasão traz numerosas perdas para os estudantes, como tempo e dinheiro, além de demonstrar o perfil das suas possíveis decisões quanto às outras formações e projetos individuais. Sendo assim, a compreensão de tal fenômeno se torna fundamental para gestão de uma instituição de ensino superior, visto que o mesmo representa um desperdício de recursos e perda de futuros profissionais.

Conforme Sultana, Khan e Abbas (2017), este problema pode ser tratado, mas para isso, as instituições devem elaborar planos mais eficientes para reduzir o índice de evasão, assim, aumentando o número de alunos formados.

Dentro desse campo de combate à evasão do ensino superior, a Mineração de Dados Educacionais (MDE) coopera ao fornecer informações que contribuam sobre as decisões em torno do processo educacional (SULTANA; KHAN; ABBAS, 2017).

Entretanto, para realizar uma análise dos grandes volumes de dados acadêmicos obtidos pelas instituições, o uso de algoritmos de Aprendizado de Máquina é imprescindível para tal estudo (SILVA; ALMEIDA; RAMALHO, 2020).

Devido aos modelos de algoritmos estatísticos, o estudo de Aprendizado de Máquina é capaz de determinar previsões de eventos ocorrerem (SILVA., ALMEIDA., RAMALHO, 2020).

Portanto, esta pesquisa tem como desafio identificar os obstáculos internos e externos que junto ao tempo agravam o fenômeno de evasão escolar, por meio de uma análise dos dados sociais a qual, futuramente, pode contribuir no treinamento de um

sistema de previsão, desenvolvido a partir de algoritmos como *decision tree, random forest* e *artificial neural network*, sobre impacto do tempo na tomada de decisão dos alunos.

2 FUNDAMENTAÇÃO TEÓRICA

A mineração de dados para o estudo de fenômenos econômicos, sociais ou históricos se caracteriza pela descoberta de padrões a fim de obter informações dos mesmos. Entretanto, com a evolução dos meios de comunicação, globalização da internet e democratização do acesso à tecnologia, o volume de dados acumulados cresceu de forma exponencial, tornando a análise manual completamente inviável.

Este aumento na captura de dados é um fenômeno característico da era contemporânea, posto que, a cada aproximadamente 20 meses, dobra-se o volume de informações armazenadas em uma variedade diversa de processos (MARR, 2015). A partir disso, foram desenvolvidas técnicas e linguagens específicas para contribuir na praticidade de estudos em cima de imensos bancos de dados.

Atualmente, a mineração de grandes conjuntos de dados educacionais é capaz de detectar correlações úteis para resolução de problemas abordados em artigos científicos que envolvem a predição de fenômenos. A evasão escolar se trata de um deles, posto que através de algoritmos de *machine learning* tornou-se possível a criação de um software capaz de determinar a possibilidade de um aluno desistir do seu curso.

2.1 MACHINE LEARNING

O machine learning (ML) tem ocupado um imenso espaço na área de tecnologia da informação ao utilizar de algoritmos e técnicas computacionais em análises de dados (SMOLA; VISHWANATHAN, 2008).

Em decorrência da expansão exponencial dos bancos de dados, tudo leva a crer que uma apuração perspicaz dos mesmos se tornou fundamental para o progresso tecnológico (SMOLA; VISHWANATHAN, 2008).

Conforme defende Amorim, Barone e Mansur (2008), o ML apresenta uma contribuição fundamental em processamento de linguagem natural, motores de busca, diagnósticos médicos, bioinformática, reconhecimento de fala, reconhecimento de escrita, visão computacional, locomoção de robôs e sistemas de previsão.

Fundamentalmente, existem quatro tipos de técnicas utilizadas no treinamento dos algoritmos de ML: supervisionado, não-supervisionado, semi-supervisionado e aprendizado por reforço.

2.1.1 Aprendizado supervisionado: Compreende o provimento de exemplos com rótulos associados ao algoritmo, os quais indicam a classe do registro referente (BATISTA, 2003), desse modo, é possível identificar que em cenários específicos ocorrem output únicos. Essa categoria de aprendizado geralmente é utilizada pelos cientistas de dados para problemas de regressão ou classificação (KUMAR., KRISHNA, 2020).

O principal objetivo deste tipo de aprendizado é prever determinadas saídas para construir modelos de predições depois de um processo de aprendizado com o fornecimento de registros etiquetados (KUMAR., KRISHNA, 2020).

2.1.2 Aprendizado não-supervisionado: Não há instâncias classificadas previamente (LORENA; CARVALHO, 2003). Essas mesmas são passadas para o algoritmo sem nenhuma espécie de rotulação, portanto, cabe à máquina construir o modelo por meio de identificação de regularidades, os clusters (BATISTA, 2003).

Dentre as várias definições de algoritmos de cluster registradas, Jain et al.(1999) afirma que estes são classificados a partir dos métodos escolhidos para formar os clusters: particionais, grade, densidade e hierárquicos.

No aprendizado não-supervisionado as rotulações sobre os dados são determinadas a partir dos agrupamentos resultantes dos padrões identificados pelo algoritmo. Dentre os mais importantes, de acordo com Géron (2019), têm-se o *Clustering,* Visualização e Redução de Dimensionalidade, *Kernel* PCA, *locally-linear Embedding* e Aprendizado da Regra da Associação.

2.1.3 Aprendizado semi-supervisionado: Alguns poucos dos registros providos são rotulados e deve-se fazer o possível com os demais não classificados, entretanto, além de haver ruídos aleatórios nos dados, os rótulos não podem ser interpretados como completamente verdadeiros, uma vez que existem imprecisões sistemáticas que cabe ao algoritmo reconhecer (STUART RUSSEL; PETER NORVIG, 1998).

2.1.4 Aprendizado por reforço: o algoritmo basicamente aprende por um sistema gamificado o que inclui uma série de reforço por recompensa e punição, cabendo ao algoritmo escolher quais das ações anteriores contribuíram para alcançar seu objetivo programado (STUART RUSSEL; PETER NORVIG, 1998).

2.2 MODELOS DE APRENDIZADO DE MÁQUINA

Para a escolha da técnica de aprendizado de máquina mais eficiente sobre o estudo de evasão escolar, deve ser levado em consideração certos fatores, como a origem dos dados coletados, a quantidade da amostra disponível e a qualidade dos registros.

Neste trabalho foram sugeridas quatro técnicas de Aprendizado de Máquina com o intuito de revelar qual dessas poderia obter uma maior acurácia no modelo estatístico para previsão de evasão escolar: *Decision Tree, Random Forest* e *Neural Network.*

2.2.1 Decision Tree

A *Decision Tree* (DT) é uma técnica de *machine learning* desenvolvida por Breiman (2001), que utiliza da organização de uma árvore para definir um número de caminhos de decisão, cada qual com um resultado (GRUS, 2016).

A técnica de DT é definida a partir de um conjunto de regras cada a qual percorre o início da raiz da árvore até uma das folhas, a condição de teste do nó raiz é aplicada ao registro para então o algoritmo dirigi-lo a ramificação de acordo com o resultado, podendo este ser um outro nó interno ou um nó folha, caso seja um nó folha, a classe associada a mesma será atribuída ao registro (GRUS, 2016).

Esse modelo se caracteriza por ser de fácil entendimento, uma vez que usam um processo transparente, lidando com atributos numéricos e categóricos, além de também conseguirem classificar dados dos atributos falantes (GRUS, 2016).

2.2.2 Random Forest

Conforme Cesare Neto (2014), *Random Forest* usa um método da DT para mineração de dados, entretanto, este método possui um objetivo diferente da *Decision Tree*, posto que a construção do algoritmo busca executar a criação de várias DT a partir de subconjuntos escolhidos de forma aleatória em relação ao conjunto original estes possuem um tipo de amostragem chamado de *bootstrap*, a qual é do tipo com reposição, possibilitando assim melhor análise dos dados

2.2.3 Artificial Neural Network (ANN)

Conforme Geron (2019), as ANNs foram criadas a partir da análise da extensa rede de neurônios presentes no cérebro humano, sendo essas amplamente utilizadas para resolução de problemas complexos que envolvem classificação ou regressão. Entretanto, devido à gradual mudança em relação à definição das ANNs, muitos estudiosos já descartam a comparação dessas células artificiais com as biológicas (Geron, 2019).

3. METODOLOGIA DE PESQUISA

Dados fornecidos pela Universidade Jorge Amado e pelo IFRO campus Ji-Paraná foram submetidos a uma profunda análise manual com o objetivo de determinar as características mais incisivas na decisão dos alunos. Desde já, se tornaria possível construir um software de previsão de evasão, esse que apresentaria uma interface com a porcentagem de chance de o aluno evadir do seu curso de origem.

Portanto, seria viável tratar esses casos com antecedência e, assim, reduzir o número de desistências dos cursos de graduação oferecidos pela instituição.

A realização de uma análise dos dados sociais fornecidos foi executada por meio de um computador virtual disponibilizado no COLAB.

4. EVASÃO ESCOLAR

Completar a formação do ensino superior se trata de um desejo de muitos indivíduos de diferentes faixas etária, uma vez que tal conquista representa uma possibilidade de mudança de vida e ascensão social e financeira.

Com a chegada à universidade, o aluno traz consigo objetivos e compromissos pré-definidos que variam de acordo com suas origens, todavia, ao decorrer do tempo, o mesmo passa por muitas experiências no ambiente acadêmico que o induz a redefinir suas intenções, podendo resultar em evadir-se (WAGNER BANDEIRA ANDRIOLA; CRISTIANY GOMES ANDRIOLA; CRISTIANE PASCOAL MOURA, 2006).

A priori, dentre os fatores que mais contribuem para a evasão do ensino superior por parte dos alunos no primeiro ano do curso, tem-se a desmotivação a qual se apresenta com maior influência na tomada de decisão dos graduandos os quais ainda não possuem um vínculo forte com a instituição em que está matriculado, de acordo com Tabak (2002) e Silva Filho (2007).

5. RESULTADOS E DISCUSSÕES

Nesta pesquisa, a definição de evasão se trata de evasão de cursos graduação que de acordo com a Comissão Especial de Estudos da Evasão do MEC (1997) descreve como a saída do aluno do seu curso de origem, sem realizar a conclusão deste, o que inclui: **ingresso em outro curso regular, matrícula cancelada a pedido, trancamentos excedidos, abandono, remanejamento interno, desligamento de ingressante, não**

renovação de matrícula, integralização excedida por projeção, integralização excedida, óbito e transferência para outra IES.

As duas fontes de registros utilizadas neste trabalho tiveram origem a partir de questionários internos realizados pela UNIJORGE e IFRO campus Ji-Paraná sobre ingressantes de diferentes anos.

Os dados socioeconômicos utilizados na dissertação de mestrado em assessoria de administração do ISCAP disponibilizado pelo IFRO são referentes apenas aos calouros dos anos 2014, 2015, 2016 e 2017, período este anterior à pandemia do *COVID-19*.

Por outro lado, os questionários dos registros fornecidos diretamente da UNIJORGE foram realizados em 2020, 2021 e 2022 sobre alunos que declararam intenção de evadir.

Figura 1 – Colunas de dados da IFRO

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 959 entries, 0 to 958
Data columns (total 27 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   CURSO                        959 non-null    object
 1   TURNO_ULTIMA_TURMA           959 non-null    object
 2   SERIE_ULTIMA_TURMA           959 non-null    object
 3   MUDOU_TURNO?                 959 non-null    object
 4   IDADE_MAT                    959 non-null    int64
 5   TEM_PAI?                     959 non-null    object
 6   SEXO                         959 non-null    object
 7   1_COR_RACA                   959 non-null    object
 8   2_RELIGIAO_CRENCA            959 non-null    object
 9   3_RESP_FINANC                959 non-null    object
 10  4_RENDA_FAMILIAR             959 non-null    object
 11  5_EXERCE_ATIV_REMU           959 non-null    object
 12  7_RECEBE_AUX_GOV             959 non-null    object
 13  9_SITUACAO_IMOVEL            959 non-null    object
 14  10_LOCOMOCAO_ATE_IFRO        948 non-null    object
 15  11_DISTANCIA_RESIDENCIA      959 non-null    object
 16  12_ASSIST_CAED?              959 non-null    object
 17  13_TIPO_ESCOLA_ANTES_IFRO    959 non-null    object
 18  14_FORMA_AQUIS_MAT_ESC       946 non-null    object
 19  15_FORMA_AQUIS_LIVRO_DID     937 non-null    object
 20  16_FORMA_AQUIS_UNIFORME      950 non-null    object
 21  17_HABILIDADE_ARTISTICA      959 non-null    object
 22  18_PRATICA_ESPORTE           959 non-null    object
 23  19_DIFICULDADE_APRENDIZAGEM  959 non-null    object
 24  20_NECESSIDADE_ESPECIAL      959 non-null    object
 25  22_ACESSO_INTERNET           959 non-null    object
 26  EVADIDO?                     959 non-null    object
dtypes: int64(1), object(26)
memory usage: 202.4+ KB
```

Fonte: Autores (2023)

Figura 2 – Colunas de dados da UNIJORGE

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2691 entries, 0 to 2690
Data columns (total 24 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   ALUNO                     2691 non-null   int64
 1   MODALIDADE                2691 non-null   object
 2   CAMPUS                    2691 non-null   object
 3   NOME_CURSO                2691 non-null   object
 4   TURNO                     2691 non-null   object
 5   SERVICO                   2687 non-null   object
 6   STATUS_SERVICO            2672 non-null   object
 7   SITUACAO_ATUAL            2691 non-null   object
 8   BENEFICIO                 2691 non-null   object
 9   ID_ANO_BASE               2687 non-null   float64
 10  FAIXA_DIVIDA              2672 non-null   object
 11  SERIE                     2687 non-null   float64
 12  TIPO_INGRESSO             2687 non-null   object
 13  ANO_INGRESSO              2687 non-null   float64
 14  DESSEMESTRALIZADO         2687 non-null   object
 15  MEDIA_FINAL_SEM_ANTERIOR  2687 non-null   float64
 16  SITUACAO_FINANCEIRA       2687 non-null   object
 17  DISCIPLINAS_A_CURSAR      2672 non-null   float64
 18  PERCENTUAL_CH_CUMPRIDA    2672 non-null   float64
 19  DATA_ATUALIZACAO          2687 non-null   datetime64[ns]
 20  NOME_CONCURSO             2687 non-null   object
 21  TIPO_CURSO                2687 non-null   object
 22  PARAMETRO                 2463 non-null   object
 23  RISCO                     2672 non-null   object
dtypes: datetime64[ns](1), float64(6), int64(1), object(16)
memory usage: 504.7+ KB
```

Fonte: Autores (2023)

Primeiramente, para análise dos registros fornecidos foi necessária a limpeza e conversão numérica do conjunto de dados, logo após, foi realizada uma investigação sobre a conexão entre os atributos e notou-se que a característica mais relevante para decisão do aluno sobre evadir se tratava do semestre a qual cursava.

A partir da apuração dos dados da dissertação de mestrado realizada na ISCAP, foi notória a correlação de estudantes do último e penúltimo semestre representava uma porcentagem baixíssima em comparação com os semestres anteriores, pois estes se tratavam apenas, respectivamente, 2.9% e 0.3% dos registros de evasão fornecidos.

6. CONCLUSÃO

Portanto, a partir do estudo realizado, foi evidenciado a importância da mineração de dados educacionais e como o uso de algoritmos de aprendizado de máquina no auxílio para identificação de padrões determinantes sobre a evasão escolar

pode se tornar uma peça-chave na criação de estratégias de combate a este fenômeno universitário.

Em síntese, este trabalho contribuiu significativamente para a discussão e compreensão da evasão escolar, ressaltando a importância da aplicação de técnicas avançadas de análise de dados na resolução de questões educacionais.

Ao analisar as correlações entre as variáveis disponibilizadas e o fenômeno da evasão, identificou-se que as colunas de maior relevância para a construção de um algoritmo especialista de previsão de evasão escolar são aquelas relacionadas ao impacto do tempo.

O tratamento e análise dos dados fornecidos confirmaram que o ano de ingresso do aluno, o semestre matriculado e o percentual de carga horária cumprida apresentam correlações significativas. Essas variáveis, portanto, são aptas a integrar o treinamento de *machine learning*, fortalecendo a precisão do modelo.

Por fim, ao unir esforços com diversas fontes e instituições de ensino superior na formação de uma rede de pesquisa e mineração de dados educacionais disponíveis para análise pública, seria possível desenvolver um software de previsão de evasão com uma acurácia consideravelmente superior.

Isso se torna ainda mais relevante, visto que, apesar de este trabalho ter apresentado resultados satisfatórios na identificação das categorias essenciais para a previsão, o banco de dados utilizado revelou-se relativamente pequeno, limitando a construção de um algoritmo ideal.

7. REFERÊNCIAS

GÉRON, Aurélien. Mãos à obra: Aprendizado de máquina com Scikit-learn & TensorFlow – conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. 1. ed. Tradução de Rafael Contatori de: Hands-on machine learning with Scikit-learn & TensorFlow. Rio de Janeiro: Atlas Books, 2019. ISBN: 978-85-508-0381-4**.**

KUMAR, V. Uday; KRISHNA, Azmira. Advanced prediction of performance of a student in an university using machine learning techniques. In: INTERNATIONAL CONFERENCE ON ELECTRONICS AND SUSTAINABLE COMMUNICATION SYSTEMS (ICESC). Agosto de 2020. Anais [...]. [S.l], agosto de 2020. DOI. 10.1109/ICESC48915.2020.9155557.

MARR, B. Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance. [S.l.]: John Wiley & Sons, 2015.

NETO, Cesare Di Girolamo. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. INPE, São José dos Campos, 2014.

SILVA, Fernanda Cristina da; CABRAL, Thiago Luiz de Oliveira; PACHECO, Andressa Sasaki Vasques. Evasão ou permanência? Modelos preditivos para a gestão do ensino superior. AAPE – Arquivos Analíticos de Políticas Educativas, [s.l.], v. 28, n. 149, 2020.

SULTANA, Sara; KHAN, Sharifullah; ABBAS, Muhammad A. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. International Journal of Electrical Engineering Education, [s.l.], v. 54, p. 105-118, 2017. DOI: 10.1177/0020720916688484.

SILVA, Andréa Ferreira da; ALMEIDA, Aléssio Tony Cavalcanti de; RAMALHO, Hilton Martins de Brito. Predição do risco de reprovação no ensino superior usando algoritmos de machine learning. Teoria e Prática em Administração, [s.l.], v. 10, n. 2, p. 58-80, jul-dez., 2020. DOI. 10.21714/2238-104X2020v10i2-51124.

WAGNER BANDEIRA ANDRIOLA; CRISTIANY GOMES ANDRIOLA; CRISTIANE PASCOAL MOURA. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da Universidade Federal do Ceará. Ensaio de políticas públicas educacionais, v. 14, n. 52, setembro de 2006.

7. COMPUTADOR VIRTUAL UTILIZADO

Link da base de dados:
https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fzenodo.org%2Frecord%2F3251764