12º SIEPE
SALÃO INTERNACIONAL DE ENSINO, PESQUISA E EXTENSÃO
TEMPO VIRTUAL, INTELIGÊNCIA ARTIFICIAL
24 a 26 de novembro de 2020
TIEMPO VIRTUAL, INTELIGENCIA ARTIFICIAL
24 al 26 de noviembre de 2020
eventos.unipampa.edu.br/siepe

ESTUDO DA EVASÃO DE DISCENTES DO CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO UTILIZANDO EDUCATIONAL DATA MINING (EDM)

(Autora: Karina Casola, Orientador: Prof. Dr. Alessandro Bof de Oliveira)
Karina Casola Fernandes, discente de Ciência da Computação, Universidade Federal do Pampa, Campus Alegrete
Prof. Dr. Alessandro Bof de Oliveira, docente, Universidade Federal do Pampa.

e-mail do primeiro autor: karinacasola@alunos.unipampa.edu.br

RESUMO: A evasão é o objeto de estudo de diversas áreas e uma preocupação recorrente em Instituições Federais de Ensino Superior (IFES), pois, está associada com a perda social e de recursos de todos os envolvidos no processo de ensino. A análise de maneira ágil dos dados que as Instituições dispõem é importante para se efetivar ações preventivas para mitigar o problema. Nesse sentido, esse trabalho realizou o estudo da evasão dos discentes do curso de Ciência da Computação da Universidade Federal do Pampa (Unipampa), utilizando métodos de *Educational Data Mining* (EDM). EDM compreende as áreas de pesquisa interdisciplinares que abrangem recortes de Ciência da Computação, Estatística, Psicologia e Educação. Para a realização desse estudo foi utilizada a linguagem de programação Python e R, a biblioteca Scikit learn, e Pandas dentro do processo *Sample Explore Modify Model and Assess* (SEMMA) e *Knowledge Discovery in Databases* (KDD) para delimitar o trabalho formalmente da representação estatística da amostra de dados, e para nortear os processos da análise preditiva. Os dados foram utilizados de maneira anônima, respeitando a privacidade dos discentes. A pesquisa atuou sob dois aspectos: A análise do perfil socioeconômico dos ingressantes através dos dados do Sistema de Seleção Unificada (SiSU)/Exame Nacional do Ensino Médio (ENEM), referente aos anos de 2010 a 2018 e a análise dos discentes matriculados no primeiro ano nos componentes curriculares referentes aos dois primeiros semestres no eixo temporal de 2009 a 2018. As análises do perfil ingressante e discente foram feitas de maneira separada não tendo o cruzamento de informações. A análise socioeconômica se centrou na análise por agrupamento e estatística, enquanto a análise do perfil discente utilizou dois algoritmos de *Machine Learning* com aprendizado supervisionado (ou seja, os dados já eram rotulados nas possíveis classes que os discentes se encontrariam, como, por exemplo: Aluno Regular, Abandono, etc). Os algoritmos utilizados para a concepção de um modelo preditivo foram o *K — Nearest Neighbors* (KNN) e Rede Neural Artificial Perceptron

Multicamadas com retro propagação, respectivamente. A análise socioeconômica seguiu o processo SEMMA de maneira adaptada, tendo quatro etapas: amostragem: — seleção dos dados que comporiam a análise, de modo a evitar ruídos na amostra. Nesta etapa foi realizada a limpeza dos dados que pudessem interferir na análise e geração de gráficos e apontamentos estatísticos, como, por exemplo, atributos faltantes em quaisquer colunas, atributos com valoração diferente do padrão seguido pela tabela de dados. Foram eliminados 8 registros que estavam inconsistentes, pois não haviam valorações para as competências específicas abordadas no ENEM, atributos faltantes da nota final, indicação de forma de ingresso, UF e município de origem. Os dados correspondiam a um montante de 450 registros antes da eliminação, restando, portanto, 442 registros. A falta de padronização dos atributos na rotulagem "Forma de ingresso", denominada como sendo "Ação afirmativa", foi resolvida com a modificação para siglas comumente usadas no processo SiSU, deixando apenas as siglas A1 e V419, adotadas pela Instituição. Eliminando assim a redundância da informação. A etapa de exploração: — concepção de uma visão por agrupamentos através da produção de um algoritmo básico utilizando a biblioteca Pandas para agrupamento das informações para a seleção de determinados critérios a serem analisados na base, por exemplo: área de competência Sisu/ENEM agrupada pela situação do discente. E a análise estatística: — de modo a verificar a distribuição destes dados através da biblioteca Pandas dentro da análise estatística descritiva básica. Na análise socioeconômica foi apurado que mesmo que a evasão tenha um alto índice, existem grupos de maior risco que possuem necessidades especiais para o aprendizado, com ações afirmativas do núcleo (Área básica do ingressante — ABI, denominada A1). Durante o período analisado, todos que ingressaram como A1 abandonaram o curso e cerca de 89% dos que ingressaram com V419 (Levando-se em conta cancelamento e transferência Interna). As notas nas competências por área da avaliação ENEM/SiSU, não demonstrou ter impacto dado à análise estatística da correlação linear das notas com a permanência desse discente.

A análise do perfil discente teve basicamente três etapas: a análise exploratória, adoção de algoritmos para a concepção de modelos preditivos e avaliação desses modelos. Se entende por predição como sendo a capacidade que o modelo adotado conseguia prever a qual classe os discentes pertenciam, e consequentemente a possibilidade de evasão. Com a etapa exploratória foi possível verificar a distribuição dos discentes pela sua situação/classe, que posteriormente foi confirmado pelo modelo preditivo. A etapa de avaliação utilizou duas análises: a matriz de confusão e a validação cruzada. Os métodos supervisionados treinados obterão uma acurácia de 98,65% e 97,89%, indicando que existe um padrão que pode ser mapeado utilizando os modelos preditivos explicitados nesse trabalho. A matriz de confusão permitiu fazer as seguintes conjecturas sobre os dados, tomando por base o algoritmo que teve o melhor desempenho de mapear e prever as classes: a classe com maior quantidade de registros é a 2, ou seja, cada vez que entrar um novo registro, ele será classificado a "priori" como pertencente à 2 "abandono". Outra técnica utilizada para verificar a precisão do algoritmo foi a *Zero Rules*, ou Zero Regras. O acerto mínimo do algoritmo foi de 43,87%, que se obteve fazendo a seguinte equação: $ClasseM/TotalR \times 100$. Onde a *ClasseM* representa a classe com quantidade maior de registros e *TotalR* a quantidade total de registros. A probabilidade de alguém selecionar a classe correta, levando-se em consideração ao número de classes, ou seja, as "situações" que os discentes podem assumir é de 14,28%. Isso é bem inferior ao acerto mínimo do algoritmo de aprendizado de

máquinas, com o que se conclui que é interessante sua adoção no planejamento de políticas para inibir a evasão. Para verificar se havia diferenças estatísticas significativas entre os dois algoritmos selecionados neste estudo, foi utilizado em R a análise de *Friedman* com *Nemenyi*, obtendo um resultado de distância crítica de 0,35, demonstrando indícios de similaridade. Os desafios deste estudo foi dispor de poucos dados históricos que se pudessem fazer inferências do comportamento do discente, como, por exemplo, apontado na literatura (ROMERO et al., 2010), o registro logs do uso de sistemas educacionais, da plataforma *Modular Object-Oriented Dynamic Learning Environment* (Moodle). E a inviabilidade da extração de traços comportamentais dos discentes em redes sociais por ferir o direito à privacidade e o anonimato dos mesmos. Ainda assim os dados históricos têm menos ruídos que a realização de entrevistas, seja por parte de informações omitidas pelo entrevistado, ou por uma inclinação, ou viés do entrevistador. Com a mineração dos dados e o modelo preditivo bem sedimentado, é possível atuar em pontos focais evitando casos de evasão, ou seja, em grupos de risco analisados no estudo.

Agradecimentos: À Universidade Federal do Pampa (UNIPAMPA) a possibilidade de ampliar os horizontes através do estudo e pesquisa.
Palavras-chave: *Educational Data Mining* (EDM)*;* Aprendizado Supervisionado; Evasão.