DATA SCIENCE EM BASE ESCOLAR: UMA PERSPECTIVA ESTATÍSTICA DA EVASÃO ESCOLAR

***Celso Barreto Da Silva*[1]**
***José Vicente Cardoso Santos*[2]**
***Ana Patrícia Fontes Magalhães Mascarenhas*[3]**

RESUMO: Um importante recurso em que todos os serem humanos precisam se apoiar é a educação. Diante da afirmação, nota-se que a educação no Brasil enfrenta problemas sérios de estrutura e infraestrutura e outros investimentos e entre os problemas catalogados, existe uns mais questionados e pesquisados é a evasão escolar. Nesse cenário essa pesquisa tem por objetivo geral estabelecer correlação entre a evasão em cursos de TI com o perfil de originário dos alunos, e, por objetivos específicos a coleta de dados para alimentação de base de dados focal; o tratamento isento dos dados coletados; o treinamento do modelo de aprendizagem de máquina, e, a visualização e análise dos resultados. Avança-se já com a construção da primeira parte do texto destaca através de fundamentação histórica e coleta de dados para que tenhamos dados para a criação do modelo e o embasamento teórico necessário ao tempo em que adota-se uma metodologia lastreada em revisão de literatura específica e aplicada para este estudo que consiste na prospecção de publicados na área (artigos, monografias, teses, dissertações bem como bases de dados oficiais) com construção de texto final a partir de consultas a textos acadêmicos e documentos pertinentes à discussão. Na segunda parte da pesquisa é debatido como sobre a limpeza dos dados com a aplicação de técnicas como a ETL com foco em maximizar o resultado para o nosso modelo ao tempo em que expecta-se o tratamento da aplicação das técnicas escolhidas para a realização do treinamento de nosso modelo Machine Learning e a catalogação dos resultados e por fim analisaremos os resultados obtidos com demonstrações estatísticas e gráficas e espera-se encontrar o melhor resultado de previsão através dos dados estatísticos acompanhados pelos gráficos e observações sobre o projeto.

PALAVRAS-CHAVE: School Dropout; Machine Learning; Data Science, Evasão Escolar; Aprendizagem.

ABSTRACT: An important resource that all human beings need to rely on is education. In light of this statement, it is evident that education in Brazil faces serious problems of structure, infrastructure, and other investments, and among the cataloged issues, there is one that is more questioned and researched: school dropout. In this scenario, the general objective of this research is to establish a correlation between dropout rates in IT courses and the profile of the students, and the specific objectives include data collection for the focus database,

[1] Mestrando em Sistemas e Computação pela UNIFACS LAUREATE – profcelsobarreto@hotmail.com
[2] Doutorado em Modelagem Computacional pela Faculdade de Tecnologia SENAI-CIMATEC - prof.vicentecardoso@gmail.com
[3] Doutora em Ciências da Computação pela Universidade Federal da Bahia e mestre em Mecatrônica pela Universidade Federal da Bahia. - anapatriciamagalhaes@gmail.com

unbiased treatment of the collected data, training of the machine learning model, and visualization and analysis of the results. The first part of the research is already underway, highlighting the historical background and data collection to provide the necessary data for the creation of the model and the required theoretical basis. This is done by adopting a methodology based on a review of specific and applied literature in the field, including the exploration of published works such as articles, monographs, theses, dissertations, as well as official databases. The final text is constructed based on consultations with academic texts and relevant documents for the discussion. The second part of the research discusses data cleaning through techniques such as ETL, with a focus on maximizing the results for our model. The application of the chosen techniques for training our machine learning model and the cataloging of the results are expected. Finally, the obtained results will be analyzed using statistical and graphical demonstrations, aiming to find the best prediction result through statistical data accompanied by graphs and observations about the project.

KEYWORDS: School Dropout; Machine Learning; Data Science; School Evasion; Learning

Introdução

Este artigo aborda o sobre o desenvolvimento de um modelo computacional relacionado a evasão escolar, onde trata-se de um dos desafios que necessitam ser superados e dominado na educação brasileira, conforme defende (PÉREZ et al., 2018; MARQUES, 2020; CASANOVA et al., 2021), a evasão em instituições escolares é um tema altamente pesquisado no mundo, pois esta representa um desafio para as instituições de ensino que afeta a reputação e credibilidade entre os diversos centros de ensino. Silva, Cabral e Pacheco (2020) afirmam que a evasão causa uma perda ao estudante, pois, não consegue obter sua formação, ele perde tempo, além de perder dinheiro, e propósito individual, e para a escola ou universidade a mesma acarreta perda de eficiência (SILVA, CABRAL, PACHECO, 2020), em acordo com Biazus (2004), as pesquisas que foram realizadas sobre o tema evasão estão com foco maior ao ensino fundamental e ao ensino médio justificando este trabalho com foco no ensino superior.

A evasão de alunos é um fenômeno complexo, comum às instituições de ensino no mundo contemporâneo e necessita de atenção (BRITO, MEDEIROS, BEZERRA, 2019). O desenvolvimento da transformação digital, incluindo a implementação generalizada da internet, levou a um aumento do número de alunos no ensino superior, inclusive por meio de cursos online/ensino a distância. No entanto, este crescimento está sendo acompanhado por altas taxas de evasão (BRITO, MEDEIROS, BEZERRA, 2019). A evasão escolar é um problema encontrado nas instituições de ensino que trazem diversos prejuízos aos seus investidores, e

impõe custos a todas as partes envolvidas, sejam recursos, tempo ou dinheiro (GANSEMER-TOPF e SCHUH 2006; YU et al., 2010). Conforme defende Rigo et al. (2014), o termo evasão escolar permite diversas interpretações e é utilizado em diferentes contextos com significados ligeiramente diversos, já para Favero (2006), evasão na não continuidade ou na desistência pelo estudante.

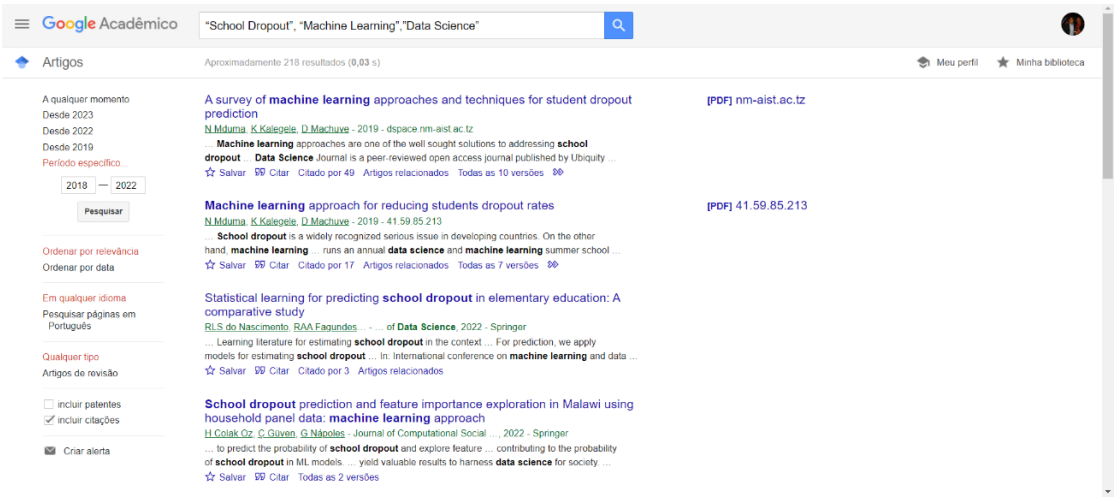
Entendendo o cenário atual da educação, será possível a realização de análises e desenvolvimento de estratégias para aumento de retenção discente, com o desenvolvimento de uma estratégia para a prevenir o abandono da escola, pois trata-se de um grande desafio às instituições de ensino (ZHANG et al., 2010). Diversas empresas tentam reduzir suas perdas de clientes realizando análises de dados e mineração de dados preditiva (DELEN, 2010). O principal objetivo deste trabalho concentra-se na análise de base de dados educacionais e o desenvolvimento de um modelo computacional com a aplicação de técnicas de Machine Learning para previsão do abandono e/ou escolar.

Este trabalho realizou uma análise e desenvolvimento de uma modelo computacional de previsão usando Machine Learning (Aprendizado de Máquina), para a realização de análises da dos dados para predição de evasão escolar. Para que esta meta fosse possível ser alcançada foi necessário estabelecer alguns objetivos específicos coma a finalidade de entendermos o cenário de maneira clara, são eles: A coleta de dados para alimentação de base de dados focal, O tratamento dos dados coletados; testar e escolher técnicas de Machine Learning com foco no melhor resultado; O treinamento do modelo de aprendizagem de máquina; A visualização e análise dos resultados.

Referencial teórico

Esta seção tem a finalidade de apresentar o referencial teórico da pesquisa e fundamentar as contribuições para as definições encontradas neste trabalho. Assim, foram encontradas no campo do conhecimento as principais referências na definição do conceito, dimensões e elementos, bem como a aplicação de modelos ou tipologias relacionados à data Science e assuntos relacionados ao tema proposto neste trabalho. Para a realização deste trabalho, foi necessário analisar alguns trabalhos anteriores sobre o assunto evasão e temas relacionados, e através deste desenvolver uma discussão da literatura. Foi desenvolvida uma pesquisa bibliométrica em abril de 2022, utilizando as palavras-chave: “School Dropout”, “Machine Learning”,”Data Science”, utilizando o portal do google acadêmico, onde foram encontrados 218 resultados, sendo que os mesmos foram filtrados num período de 2018 a 2022, conforme imagem abaixo.

Figura 1-Tela de busca do portal google acadêmico



Fonte: Pesquisa realizada pelo autor.3

Após análise dos trabalhos não relacionando e repetidos, e alguns que não estavam disponíveis totalmente, restaram 73 trabalhos com níveis de relevância alta para contribuir com este artigo.

Na sequência em desenvolvimento da leitura dos trabalhos, ainda fora excluídos outros trabalhos que o seu foco não era de análise de evasão escolar restando 25 trabalhos para serem utilizados com foco em análise de dados, evasão e Machine Learning.

Evasão escolar - School Dropout

A evasão escolar, também conhecida como abandono escolar, é defendido por (TINTO 1982) como um problema complexo, e está ligada a diferentes tipos de abandono e seus comportamentos que o envolvem. É claro que a definição de abandono requer a análise de uma perspectiva que é adotado no trabalho. Existe interesse na comunidade científica por definições e pesquisas que definam modelos conceituais adequados para evasão escolar, com a finalidade de entender, descrever, caracterizar e prever este fenômeno (TINTO, 1982). Ainda sobre o tema evasão escolar, Baggi e Lopes (2011, p. 356) afirmam:

> É um problema que vem preocupando as instituições de ensino em geral, sejam elas públicas ou particulares, pois a saída de discentes provoca graves consequências sociais, acadêmicas e econômicas (Baggi e Lopes, 2011, p. 356).

[3] Disponível em: https://scholar.google.com.br. Acesso em ago. 2022.

O Ministério da Educação (MEC) define da seguinte forma o termo evasão ou abandono Lüscher (2011, p.158),

> No censo escolar do Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP), do Ministério da Educação (MEC), a saída de estudantes da escola é conceituada como abandono: refere-se apenas ao estudante que deixou de frequentar uma determinada escola em um dado ano (Lüscher, 2011, p.158).

Para este, é considerado evasão escolar a saída de maneira antecipada de um discente ou antes da conclusão do ano, série ou ciclo, por desistência (independentemente do motivo). Ou seja, quando os alunos iniciam seus cursos, mas em algum momento antes de se tornar concluinte, os cessa (BOMFIM, 2021. p.19).

Aprendizado de Máquina - Machine Learning

O temo Machine Learning ou aprendizado de máquina trata-se de um processo que através dele é possível gerar novos conhecimentos, com análises realizadas em uma base de dados (Alpaydin, 2010). Através desta inferência em bases de dados ou mineração de dados, (Witten & Frank, 2005) defende que a mesma utiliza o aprendizado de máquina na análise e obtenção de padrões existentes nos dados. Conforme (FAYYAD, Usama, 1996),

> O conceito principal de mineração de dados é a extração de conhecimento útil a partir dos dados coletados para resolver problemas de negócios, com processos sistemáticos unidos à ferramentas computacionais, evoluindo o processo entre etapas previamente definidas.

Através da mineração de dados é possivelmente realizar obtenção de conhecimentos contidos em bases de dados para poder realizar a predição de padrões. (Witten e Frank, 2005).

Ciência dos dados - Data Science

A definição de Ciência de Dados conforme (IGUAL E SEGUÍ, 2017), é uma metodologia com a qual é possível através de modelos e hipóteses inferir a partir dos dados. Pinker e Motta (2018) defende que a Ciência é um processo de refutações e algumas conjecturas, e faz com que se aumenta a confiança em uma hipótese à medida que as provas se acumulam. Portanto, o processo de realização de Ciência de Dados, serve para produzir conjecturas e refutações a partir dos dados, para que sejam utilizados de maneira assertiva no processo de tomada de decisão. Portanto, a criação, modelagem e representação de ambientes complexos com base nos dados,

exibe inúmeras possibilidades de aplicação de vários conhecimentos científicos com o objetivo de inferir modelos sobre eles (IGUAL e SEGUÍ, 2017).

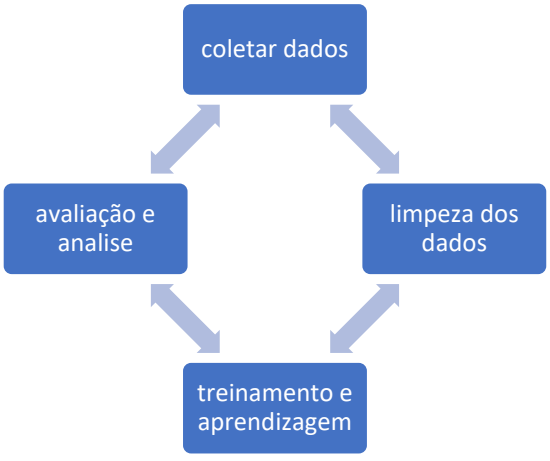
Preparação dos dados

Ao iniciar um processo de preparação e análise dos dados, é necessário antes a realização de alguns passos para a correção de problemas comuns, entre eles destacamos: a) presença de dados ausentes, b) dados discrepantes, c) dados inconsistentes, d) dados redundantes. Para que o trabalho seja realizado com alto grau de qualidade é necessário a detecção de maneira antecipada dos problemas e realizar o devido tratamento. Conforme destaca Faceli et al. (2011), existem casos que dependendo do problema existira a necessidade de exclusão dos dados problemáticos. Nesse caso, para manter-se a integridade e qualidade dos dados, realizou-se a limpeza da base de dados.

Coleta dos dados

A população para a realização deste trabalho será os dados acadêmicos dos alunos do curso de computação do centro universitário Jorge Amado. Vale registrar de conforme a Lei Geral de Proteção de Dados Pessoais (LGPD ou LGPDP), Lei nº 13.709/2018 que regula o tratamento e análises de dados pessoais, aos dados aqui analisados não serão expostos e não haverá nenhuma citação ou identificação pessoal de quaisquer alunos da instituição. Para que seja realizado o processamento e análise, faz-se necessário a preparação dos dados brutos, as principais etapas consistem na coleta, limpeza dos dados para assim adequar aos algoritmos de Machine Learning (ML), seguindo com o treinamento e avaliação e análise do modelo.

Figura 2 – Ciclo da informação



Fonte: Elaborado pelo autor (2022).

Limpeza de dados

Na etapa de limpeza dos dados, foram executadas ações de análise preliminar, ou seja, foram verificados atributos em banco, inválidos e demais atributos não condizentes com o escopo proposto nesta pesquisa.

No intuito de garantir privacidades dos dados acadêmicos conforme a Lei Geral de Proteção de Dados Pessoais (LGPD ou LGPDP), Lei nº 13.709/2018 que regula o tratamento e análises de dados pessoais os campos com informações dos nomes dos alunos e suas respectivas matrículas foram removidas da pesquisa. Demais atributos sem contribuição para a pesquisa foram removidos para criar uma base de dados enxuta e com dados relevantes, são eles: modalidade, campus, status_servico, id_ano_base, faixa_divida, dessemestralizado, disciplinas_a_cursar, percentual_ch_cumprida, data_atualizacao, nome_concurso, tipo_curso, risco.

Para a realização da análise na etapa de limpeza de dados restaram os atributos, NOME_CURSO, TURNO, SERVICO, SITUACAO_ATUAL, BENEFICIO, SERIE, ANO_INGRESSO, TIPO_INGRESSO, MEDIA_FINAL_SEM_ANTERIOR, SITUACAO_FINANCEIRA, PARAMETRO.

Treinamento do modelo e aprendizagem

Para a realização desta etapa foi utilizado a linguagem Python, pois conforme (MARQUES, 2020. p.39),

> É uma linguagem de programação de uso geral que pode ser usada para diferentes fins: coleta de dados, engenharia de dados, análise, Web Scraping, construção de aplicativos web e etc (MARQUES, 2020. p.39),

Métricas de avaliação

O modelo de classificação de dados realiza predição baseando-se em ocorrências passadas. É necessário a utilização de dados de entrada entradas que neste caso são as pessoas ou indivíduos e seus atributos, mas além disto precisa-se conhecer antecipadamente o resultado que é esperado. Para treinar um modelo é necessário utilizar todas estas informações com foco na predição dos resultados esperados e para novos dados que possivelmente irão surgir futuramente. Para a realização do modelo é necessário utilizar dados que não foram utilizados no momento do treinamento do mesmo para análise a margem de acerto do modelo. Com isto,

observa-se que não basta a realização da contagem de quantidades de acertos para informar que o modelo atingiu o seu objetivo, primeiro analisa-se o problema para aplicação de métricas diferentes.

Para avaliar um modelo de aprendizagem de máquina, algumas métricas são comumente adotadas. As métricas utilizadas serão:" Acurácia","Precisão" e "Revocação", ambas com base na matriz de confusão, e a partir delas, o F-measure.

Figura 3. Matriz de confusão

		Classe Predita	
		Evadidos	Não Evadidos
Classe Real	Evadidos	Falso Positivo (FP)	Verdadeiro Negativo (VN)
	Não Evadidos	Verdadeiro Positivo (VP)	Falso Negativo (FN)

Fonte: Elaborado pelo autor (2022).

Verdadeiros Positivos: É classificação correta da classe Positivo;
Falsos Negativos (Erro Tipo II): Previsão encontrada pelo modelo quando o valor real deveria ser Positivo;
Falsos Positivos (Erro Tipo I): Previsão encontrada pelo modelo quando o valor real deveria ser Negativo;
Verdadeiros Negativos: É classificação correta da classe Negativo.

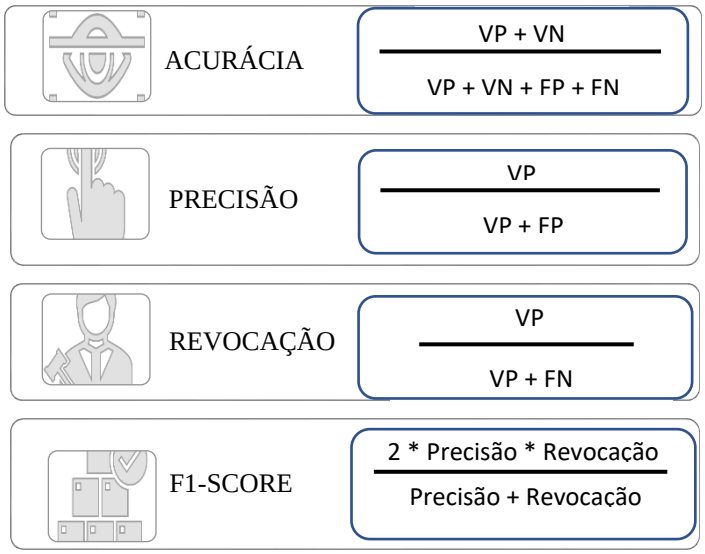
A Classe Real é a classe dos dados de entrada. O Verdadeiro Negativo (VN) diz respeito a um aluno que não evadiu e que o modelo apontou como não evasão.

O Falso Negativo (FN) são casos em que o modelo aponta não evasão, mas correspondem a um aluno que evadiu.

O Falso Positivo (FP) corresponde que o modelo apontou como evasão, mas não houve. Por fim, o Verdadeiro Positivo (VP) trata-se do aluno em que o modelo apontou como evasão, e ele evadiu.

Após realizar a contagem dos termos e montar a matriz de confusão, é possível calcular métricas de avaliação para a classificação conforme imagem abaixo:

Figura 4. Métricas de avaliação



Fonte: Elaborado pelo autor (2022).

A equação1 corresponde a acurácia, que se trata de é uma razão entre o total de predição e o número de predições corretas. A equação no quadro 2, é a razão dos VP sobre os VP e os FP, ou seja, é uma medida que está relacionada a quantos do total de valores que o modelo apontou como evasão (VP + FP) são casos de evasão de fato, estamos falando da precisão. O quadro 3, descreve a equação do Recall ou Revocação, que é a razão dos VP sobre todos os reais casos de evasão (VP e FN). A equação 4, está of1-score ou F-measure, e algumas literaturas também o chama de F1, que é a média harmônica entre a precisão e o recall.

Aplicação

Neste trabalho, foi utilizado a linguagem Python para a realização da análise e validação do modelo, assim como a manipulação de dados. Juntamente com a linguagem, foi adicionado ferramentas para a visualização e validação dos resultados que são as bibliotecas NumPy, Pandas e Matplotlib. Os scripts foram desenvolvidos em ambiente Google Colab.

Conforme COELHO (2017),

> O Python é uma linguagem de programação dinâmica e orientada a objetos, que pode ser utilizada no desenvolvimento de qualquer tipo de aplicação, científica ou não. O Python oferece suporte à integração com outras linguagens e ferramentas, e é distribuído com uma vasta biblioteca padrão. Além disso, a linguagem possui uma sintaxe simples e clara, podendo ser

aprendida em poucos dias. O uso do Python é frequentemente associado com grandes ganhos de produtividade e ainda, com a produção de programas de alta qualidade e de fácil manutenção.

Carregamento dos dados

Os dados foram carregados em ambiente google colab com a finalidade de serem manipulados e analisados posteriormente. O ambiente google Colab conforme defende (M. Canesche et.all. 2021), ”O Google Colab é um notebook Jupyter em nuvem amplamente usado para ensinar aprendizado de máquina escrevendo explicações de texto e códigos Python por meio do navegador”.

Para isto foi executando o script:

- import pandas as pd, import numpy as np

A finalidade deste script é importar para o ambiente de análise e manipulação as bibliotecas pandas e numpy.

Para o carregamento e visualização dos dados preliminares foram executadas as seguintes rotinas.

- data = pd.read_excel('Base_20.1_a_22.1.xlsx')
- data = pd.DataFrame(data)
- data

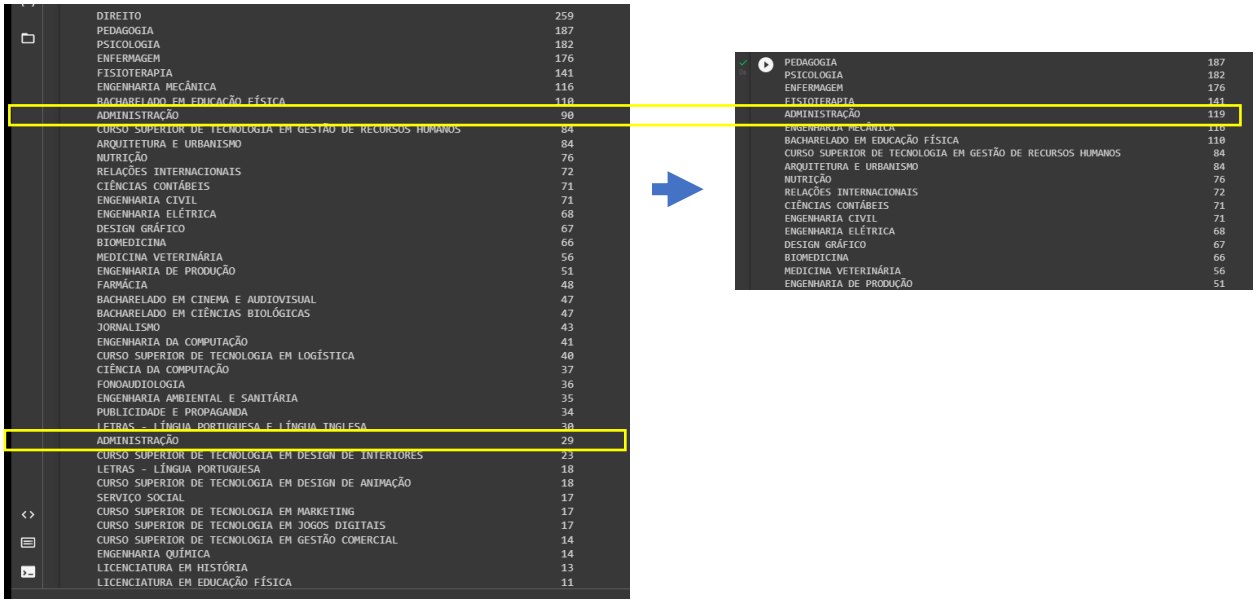
Alguns dados foram removidos por não apresentar relevância ao trabalho e outros para garantir a confidencialidade dos registros contidos na base de dados conforme Lei nº 13.709/2018. Foram removidos os dados observados no tópico 3.2 limpeza de dados e em seguida foi realizada a contagem das variáveis restantes e foram encontradas 11, para a realização de manipulações e análises. Foi utilizado o seguinte script para contagem:

print ('A base de dados apresenta {} registros e {} variáveis'.format(data.shape[0], data.shape[1])).

Após execução do script se obteve o seguinte resultado:

A base de dados apresenta 2691 registros e 11 variáveis

Figura 5. Contagem dos cursos na base de dados



Fonte: Elaborado pelo autor (2022).

Analisando os dados

Foi realizada uma contagem inicial na base, com a finalidade de encontrar o quantitativo de cursos envolvidos na análise e confrontado com o campo evadido, o que nos revelou inicialmente diversas informações. Inicialmente, foi percebido que o curso Administração está apontando um quantitativo de 90 registros e em outra linha 29 registros, podendo causar uma anomalia e erro nos resultados. Foi verificado que, o nome do curso estava com espaço no final da digitação enquanto o segundo estava digitado corretamente. Foi realizada a devida manutenção para a retirada do espaço no final dos 90 registros. Ao confrontar os campos NOME_CURSO com o atributo SERVICO, se observou que o ocorrem com frequência o preenchimento inadequado das variáveis. O campo serviço exibe 178 ocorrências com preenchimento de apenas um traço no local, causando um erro grave de análise. Foi necessário a observação do campo SITUACAO_FINANCEIRA para poder inferir acerca das informações contidas no mesmo. Obteve-se 123 registro com onde o aluno apresentava a situação ADIMPLENTE, e destes 83 alunos apresentam o n o campo SITUACAO_ATUAL como veterano, o que leva a crer que não houve evasão.

Encontramos a evasão por curso registrado na base no período de 2020-1 á 2022-1, após o analisar o quantitativo de ocorrência contidas na base para o curso selecionado em confronto com o atributo serviço, conforme tabela a seguir.

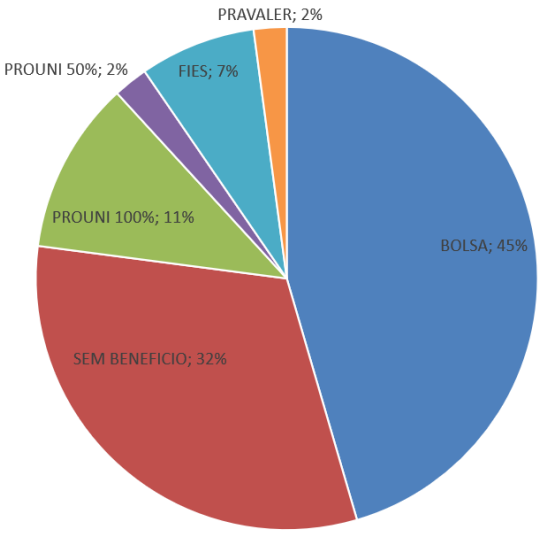
Tabela 1 – Análise da evasão por curso e serviço solicitado

Curso Graduação	Quantidade		Total
	Ocorrências	Representação	
DIREITO	259,00	20,08%	259,00
PEDAGOGIA	187,00	14,50%	187,00
PSICOLOGIA	182,00	14,11%	182,00
ENFERMAGEM	176,00	13,64%	176,00
FISIOTERAPIA	141,00	10,93%	141,00
ADMINISTRAÇÃO	119,00	9,22%	119,00
ENGENHARIA MECÂNICA	116,00	8,99%	116,00
BACHARELADO EM EDUCAÇÃO FÍSICA	110,00	8,53%	110,00
TOTAL	1290,00	100%	1290,00

Fonte: elaborada pelo autor(2022).

Foi encontrado o seguinte, o curso DIREITO é exibido com o maior número de ocorrências, representando mais de 20% das evasões por curso, seguido por PEDAGOGIA 14.50%, PSICOLOGIA 14.11%, ENFERMAGEM 13.64%, FISIOTERAPIA 10.93%, ADMINISTRAÇÃO 9.22%, ENGENHARIA MECÂNICA 8.99%, BACHARELADO EM EDUCAÇÃO FÍSICA 8.53%. Os demais cursos apresentaram menores ocorrências e por este motivo foram deixados fora da análise para a produção deste trabalho. Com a finalidade e realizar uma análise mais aprofundada, foi realizada um confronto entre os atributos, curso, serviço e benefício para poder verificar se a evasão se dá em maiores caso com alunos que não possuem benefício publico ou incentivo por benefício concedido pela instituição.

Figura 6. Análise de evasão por tipo de benefício e curso



Fonte: elaborada pelo autor (2022).

Após a execução do script para obtenção do resultado da análise, foi exibido que quase 45% dos alunos evadidos possuem BOLSA (auxílio) para a continuidade dos estudos, e em seguida encontra-se 31,59% dos evadidos contidos na base não usufruíam de algum benefício, seja ele concedido pela verba publica ou concedido pela instituição de ensino. Os demais dados obtidos da análise foram, PROUNI 100% com participação de evasão de 11,11%, seguido de PROUNI 50% com de evasão de 2,23%, FIES representa 7,47% da evasão contida na base analisada e o benefício PRAVALER continha uma evasão de 2,12%, conforme demostrado no gráfico 1 e na tabela 2 logo abaixo.

Tabela 2 – Análise da evasão por curso e benefício

BENEFÍCIO	Representação	Total
BOLSA	45,48%	259,00
SEM BENEFICIO	31,59%	187,00
PROUNI 100%	11,11%	182,00
PROUNI 50%	2,23%	176,00
FIES	7,47%	141,00
PRAVALER	2,12%	119,00
TOTAL	100%	1290,00

Fonte: elaborada pelo autor(2022).

Outra análise realizada base de dados foi a condição de cada aluno ao solicitar o cancelamento, ou seja, se o aluno é CALOURO ou VETERANO, se o discente se encontra inadimplente ou adimplente, e a sua quantidade de disciplinas que tinha disponível para cursar.

Ao executar o algoritmo, foi encontrado que entre os alunos registrados com situação de evasão, 63,23% deles estão adimplentes com a faculdade e 36,77% encontram-se com dívidas abertas e não resolvidas, que podem ser tratadas como fatores decisivos para o processo decisório em evadir da instituição. Ao aplicar nova análise em busca de um novo olhar sobre o fator decisório para evadir, foi encontrado que entre os alunos calouros adimplentes representam 36,58% enquanto inadimplentes 39,78% e os alunos veteranos adimplentes somam 63,42% para um total de 60,22% dos inadimplentes. Estes dados sendo analisados de maneira lógica nos mostra um caminho claro de evasão em maior número entre os alunos veteranos pois os mesmos somam um total após análise final de 62,24% entre alunos evadidos para um total de alunos calouros de 37,76%.

Resultado e discussão

A partir dos resultados obtidos através das análises realizadas na base de dados adquirida, foi possível identificar que os cursos (DIREITO, PEDAGOGIA, PSICOLOGIA, ENFERMAGEM, FISIOTERAPIA, ADMINISTRAÇÃO, ENGENHARIA MECÂNICA, BACHARELADO EM EDUCAÇÃO FÍSICA) são aos cursos da instituição que mais apresentou evasões no período analisado. Foi possível encontrar que entre estes discente evadidos 45,48% possuem bolsa parcial ou integral, mas não impediu o aluno de evadir. Entre os alunos veteranos e calouros foi verificado que a maior taxa de evasão se encontra entre alunos veteranos que corresponde a 62,24%, ou seja, alunos da casa tem maior possibilidade de evadir dos cursos que alunos novos. Nesta análise foi desconsiderado o campo nota final, pois não apresentava nenhuma relação com os resultados anteriores e impactos relevantes na pesquisa.

Conclusão

Este trabalho foi desenvolvido baseando-se em base de dados, os históricos acadêmicos de 2691 alunos dos cursos de graduação, com matriculas ativas na instituição entre 2020/1 a 2022/1. Os resultados obtidos, com a aplicação de métricas de avaliação de desempenho, mostraram que é totalmente possível realização a previsão dos alunos com risco de evasão dos cursos de graduação. Tendo em vista, o contexto da base de dados fornecida, foi necessário estabelecer critérios, como uma possibilidade de evasão, então foi classificado com base nos critérios de nosso script os alunos como evadidos e não evadidos. O estudo também permitiu aplicação de data Science com a finalidade de identificar os métodos utilizados ao problema da evasão escolar. Tendo como base presente estudo, e dada a importância do problema da evasão encontrado nas instituições de ensino, frente ao crescente interesse pelo tema, bem como a carência por trabalhos correlatos, tanto nacionais e internacionais considera-se finalmente que o presente trabalho contribui para a produção científica, o que justifica sua relevância acadêmica.

Referências bibliográgicas

ALPAYDIN, E. (2010). **Introduction to Machine Learning**. The MIT Press. [GS Search]

Baepler, P., & Murdoch, C. J. (2007). Academic analytics and data mining in higher education. ***International Journal for the Scholarship of Teaching and Learning***, 4(2). Disponível em: <http://dx.doi.org/10.20429/ijsotl.2010.040217>. Acesso em 09 set. 2022.

BAGGI, Cristiane Aparecida Dos Santos e Lopes, Doraci Alves. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: **Revista da Avaliação da Educação Superior** (Campinas) [online]. 2011, v. 16, n. 2 [Acessado 09 setembro 2022], pp. 355-374. Disponível em: <https://doi.org/10.1590/S1414-40772011000200007>. Epub 28 Jun 2011. ISSN 1982-5765. https://doi.org/10.1590/S1414-40772011000200007.

BOMFIM, Glédison de Avila. **EVASÃO ESCOLAR EM INSTITUIÇÃO PRIVADA DE ENSINO SUPERIOR: ANÁLISE E PREDIÇÃO**. Disponível em: < https://www.unifacvest.edu.br/assets/uploads/files/arquivos/0e944-bomfim_gledison_avila-evasao-escolar-em-instituicao-privada-de-ensino-superior_tcc_ii_unifacvest_2021.pdf> Acesso em: 09/09/2022

BRASIL. LEGP – **Lei Geral de Proteção dos dados**. Disponível em <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 09/09/2022.

BRITO, M.; Medeiros, F.; Bezerra, EP. **Uma ferramenta baseada em infográficos para monitorar o risco de evasão no ensino a distância no ensino superior**. Disponível em:<https://ieeexplore.ieee.org/document/8937361>. Acesso em: 09/09/2022.

COELHO, Flavio Codeço. **Computação científica com Python. Uma introdução à programação para cientistas**. Petrópolis RJ. Edição do Autor. 2007. p.28. Disponível em: <http://complex.if.uff.br/_media/python_flavio.pdf>. Acesso em: Acesso em: 09/09/2022.

DELEN, D. 2010. “Uma análise comparativa de técnicas de aprendizado de máquina para retenção de alunos Gestão." **Sistemas de Apoio à Decisão** 49 (4): 498–506.

GANSEMER-TOPF, AM e JH Schuh. 2006. “Seletividade Institucional e Gastos Institucionais: Examinando Fatores Organizacionais que Contribuem para Retenção e Graduação”. **Research in Higher Education** 47 (6): 613–642.

LÜSCHER, Ana Zuleima; DORE, Rosemary. Política educacional no Brasil: educação técnica e abandono escolar. Revista Brasileira de Pós-Graduação. Brasília, supl. 1, v. 8, p. 147 - 176, 2011.

M. Canesche, L. Bragança, O. P. V. Neto, J. A. Nacif and R. Ferreira, "Google Colab CAD4U: Hands-On Cloud Laboratories for Digital Design," 2021 IEEE **International Symposium on Circuits and Systems (ISCAS)**, Daegu, Korea, 2021, pp. 1-5, doi: 10.1109/ISCAS51556.2021.9401151.

MARQUES, Leoanardo Torres. **MATEO**: UMA ABORDAGEM DE DESCOBERTA DE CONHECIMENTO PARA DESVENDAR AS CAUSAS DA EVASÃO ESCOLAR. Disponível em: <

https://repositorio.ufersa.edu.br/bitstream/prefix/5425/1/LeonardoTM_DISSERT.pdf> Acesso em: 09/09/2022.

MOHER, D. et al. The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. 2009. Disponível em: <www.prisma-statement.org> Acesso em: 09/09/2022.

TINTO, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. **Review of Educational Research**, 45(1), 89–125. Disponível em: <https://doi.org/10.3102/00346543045001089>. Acesso em: 09/09/2022.

WITTEN, I. H., & Frank, E. (2005). **Practical machine learning tools and techniques**. Morgan Kaufmann Series in Data Management Systems. Disponível em: < https://scholar.google.com.br/scholar?hl=pt-BR&as_sdt=1%2C5&as_vis=1&q=Practical+machine+learning+tools+and+techniques&btnG=>. Acesso em: 09/09/2022.

ZHANG, Y., S. Oussena, T. Clark e H. Kim. “Usando a mineração de dados para melhorar a retenção de alunos no ensino superior: um estudo de caso”. **Conferência Internacional sobre Sistemas de Informação Empresarial**. 2010.