União de Dados por Clusterização para Construção de Modelos de Predição de Evasão

Cledjan Torres da Costa, UFPI, cledjan@ufpi.edu.br
https://orcid.org/0009-0003-5144-0727
Maurílio Lacerda Leonel Júnior, UFDPar, mauriliojr21@ufdpar.edu.br
https://orcid.org/0009-0008-5993-3004
André Macedo Santana, UFPI, andremacedo@ufpi.edu.br
https://orcid.org/0000-0002-0062-1806

Resumo: O rápido avanço da tecnologia tem gerado grandes volumes de dados que, por meio do processo de Descoberta de Conhecimento em Bancos de Dados, podem proporcionar significativos benefícios no cenário educacional para instituições, alunos, professores e colaboradores. Este estudo utiliza esse processo para a predição da evasão escolar, com foco em uma situação em que o conjunto de dados não é extenso individualmente, mas que, quando agrupado com outros conjuntos, pode trazer melhorias na predição. Para isso, adotou-se uma abordagem de duas etapas: uma não supervisionada, de clusterização, seguida por uma supervisionada, de classificação. Os resultados demonstram a viabilidade da adoção de agrupamentos entre cursos por meio de algoritmos de clusterização para maximizar a capacidade preditiva dos modelos.

Palavras-chave: mineração de dados educacionais, predição de evasão, clusterização, classificação

Data Union by Clustering for Creation Dropout Prediction Models

Abstract: The rapid advancement of technology has generated large volumes of data, which, through the process of Knowledge Discovery in Databases, can provide significant benefits in the educational scenario for institutions, students, teachers, and collaborators. This study utilizes this process for predicting school dropout, focusing on a situation where the dataset is not extensive individually but, when grouped with others, can bring improvements in prediction. For this purpose, a two-step approach was adopted: an unsupervised clustering step followed by a supervised classification step. The results demonstrate the feasibility of adopting groupings among courses through clustering algorithms to maximize the predictive capacity of the models.

Keywords: educational data mining, prediction dropout, clustering, classification

1. Introdução

Segundo INEP (2022) 59% dos alunos que ingressaram em seus cursos de Graduação em 2012 evadiram até 2021 (39% até o final do 3º ano de ingresso). Na instituição estudada, Universidade Federal do Delta do Parnaíba - UFDPar, esses números chegam a ser bem mais altos, como em seus cursos de Engenharia de Pesca e Matemática, que apresentam Taxas de Desistência Acumulada até 2021 de 75% e 79,5%, respectivamente, no mesmo coorte (ingressantes em 2012).

A evasão é um problema complexo que atinge diretamente os estudantes e compromete todo o contexto social e econômico em que estão inseridos (Dutra et al., 2022). Para combater esse problema, a Mineração de Dados Educacionais (MDE) emergiu como uma forte ferramenta, automatizando o processo de análise de dados. A MDE é uma área que explora estatística, aprendizado de máquina e algoritmos de

mineração de dados aplicados a diferentes tipos de dados de ensino e permite descobrir novos conhecimentos, de modo a estabelecer bases para um processo de aprendizagem mais eficaz (Romero, 2010). Ela advém de uma das aplicações dadas à Mineração de Dados e se subdivide em várias linhas de pesquisa, como Predição, Agrupamento e Mineração de Relações (Baker et al., 2011).

Atualmente, há um crescimento no número de trabalhos publicados tanto no Brasil como internacionalmente sobre MDE no contexto da evasão escolar (Dos Santos 2021). Entretanto, existem diversos obstáculos que por vezes impedem melhores resultados, como a quantidade de dados disponíveis, em especial para instituições com informatização recente, que é o caso da instituição analisada neste estudo. A precisão da predição, taxa de acertos e outros elementos, depende da qualidade e da quantidade de dados de entrada (Manhães, 2020).

O objetivo deste trabalho é aprimorar modelos de predição de evasão através de uma estratégia para enriquecer os dados ao agregar informações de cursos com maior similaridade, identificada por meio de clusterização. Este estudo descreve um processo de MDE para predição de evasão sobre dados que só estão disponíveis com a estrutura necessária para análises a partir de 2012, resultando em um baixo volume de informações, sobretudo quando se avaliam os contextos dos cursos individualmente.

2. Trabalhos Relacionados

Pesquisadores têm explorado diversas técnicas e algoritmos em MDE para abordar a evasão, incluindo aprendizado de máquina, redes neurais e árvores de decisão. Isso reflete a complexidade do fenômeno da evasão e a ausência de uma abordagem padrão eficaz, dada a variação significativa das causas entre instituições educacionais. Assim, é essencial uma abordagem personalizada que considere as peculiaridades de cada ambiente educacional e busque inovações para enfrentar esse desafio.

O mapeamento sistemático da literatura conduzido por Marques et al. (2019) identifica as melhores ferramentas, técnicas e fatores indutores da evasão. Destacam-se as ferramentas Weka e Mplus, além de mencionar Python, com as bibliotecas Pandas e Scikit-Learn, como uma linguagem de programação de uso geral. A técnica mais utilizada é a Classificação, representando 27% dos casos, e a maioria dos trabalhos aborda fatores indutores ligados às características individuais dos estudantes.

No estudo de De Brito et al. (2020), o algoritmo *Random Forest* foi empregado para identificar características relevantes no contexto da evasão usando dados acadêmicos e demográficos. Ao agrupar cursos por Grande Área de conhecimento, alcançaram acurácia superior a 75% nas áreas de Engenharia e Exatas e da Terra. A característica cor/raça foi determinante em todos os cenários.

Saraiva et al. (2021) utilizaram o algoritmo de clusterização *K-Means* para tentar criar, dentro de um conjunto de dados relacionado a um curso específico, agrupamentos de atributos de modo a criar perfis de discentes propensos a evadir ou não. Ao realizar a análise descritiva dos três clusters resultantes, foi possível identificar um conjunto de indicadores relacionados com o desempenho acadêmico dos alunos, como turno de estudo, gênero do estudante, faixa etária, dentre outros.

Na tese de doutorado de Manhães et al. (2015), é criado uma arquitetura em camadas utilizando técnicas de MDE para predição de evasão ao término de cada período letivo. Nela, os estudos de caso analisados foram utilizados para avaliar 12 algoritmos de classificação, obtendo o *Naive Bayes* com melhor resultado geral.

O trabalho de Santos et al. (2021) utiliza técnicas de classificação com base no desempenho acadêmico dos alunos para prever a evasão. Eles criaram vários modelos preditivos, correspondendo ao número de períodos recomendados para a conclusão do curso. Os resultados alcançaram acurácias entre 79,31% e 98,25%, demonstrando que é possível prever a evasão com base apenas no desempenho acadêmico dos estudantes.

Em Viana et al. (2022) é realizado um processo de MDE para predizer evasão de dois cursos da mesma área. O estudo também adotou uma abordagem que divide os dados por períodos, do 1º ao 6º, gerando um modelo para cada um com um conjunto de atributos compartilhados e outro específico do período (dados acadêmicos cumulativos), considerando, assim, o momento em que o discente se encontra no curso.

Este trabalho difere dos demais pois utiliza o processo de clusterização, com a utilização do algoritmo *K-Means*, anterior a fase de classificação, objetivando o aumento de dados de forma não sintética. Ao identificar os cursos mais próximos entre si, através de suas características e de seus discentes, por meio de clusterização, novos conjuntos de dados podem ser criados a partir da união dos dados que compõem cada grupo/cluster identificado, objetivando criar modelos mais robustos na etapa de classificação. Com mais exemplos, visa-se uma maior capacidade de generalização, pois os modelos de predição têm a oportunidade de aprender padrões mais complexos e representar melhor a relação entre as características dos dados e as saídas desejadas.

3. Materiais e Métodos

A instituição alvo deste estudo foi criada em 2019 através de desmembramento de um campi da Universidade Federal do Piauí - UFPI. Seus cursos não possuem agrupamento em forma de Departamentos, Centros de Ensino ou Faculdades, como costuma acontecer em instituições maiores. Seus cursos são: Psicologia (PSI), Pedagogia (PED), Ciências Contábeis (CON), Ciências Econômicas (ECON), Administração (ADM), Turismo (TUR), Biomedicina (BIOM), Fisioterapia (FIS), Engenharia de Pesca (PES), Ciências Biológicas (BIO), Matemática (MAT) e Medicina. Este último, por ser bem mais recente, não foi avaliado. Os dados só estão disponíveis com a consistência e estruturação necessárias para estudos e análises a partir de 2012, após a implantação do sistema de gestão acadêmica utilizado pela instituição, restando um baixo volume de informações chave.

Para viabilizar a interpretação dos dados, transformando-os em informações úteis, foi aplicado neste trabalho o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD). Fayyad et al. (1996), pioneiro na área de KDD, o qual ajudou a estabelecer as bases conceituais para a disciplina, o caracteriza como um processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis em dados. É composto pelas seguintes etapas: coleta de dados, pré-processamento, transformação, mineração de dados e avaliação dos resultados, sendo possível retornar às fases anteriores, refinando-as até que o modelo se torne adequado às necessidades.

A proposta deste trabalho é composta por duas etapas: uma não supervisionada (Etapa 1), de clusterização, para identificação dos cursos mais próximos entre si; e outra supervisionada (Etapa 2), de classificação, na qual são utilizados os resultados da etapa anterior para criar conjuntos de dados robustos, gerados pela união dos dados dos cursos pertencentes a um cluster formado e conseguinte criação dos modelos de predição destes e dos cursos individualmente, os quais servirão para comparação com os modelos agrupados, utilizando-se para tal o Teste de Hipótese Z. Nessa etapa, são avaliados os 8 primeiros períodos cronológicos da vida acadêmica do discente, dada a quantidade as informações disponíveis. A Figura 1 esquematiza o processo realizado nas duas etapas.

A Coleta foi realizada em maio de 2023 e restringiu-se ao nível de graduação na modalidade presencial regular, contendo discentes com ingresso a partir de 2012.1 até 2023.1. O período da pandemia foi considerado, no qual as aulas foram realizadas de forma remota. Com a coleta, foram gerados 2 repositórios distintos: R1 para a Etapa 1 contendo todos os discentes com status ativo, concluído e evadido, atributos acadêmicos, características individuais dos discentes e características dos cursos aos quais estão vinculados, totalizando 24 atributos e 11263 registros; e R2, para a Etapa 2, contendo os

discentes com status evadido e concluído e atributos dos mesmos tipos descritos para a clusterização, totalizando 122 atributos e 5979 registros. Para o pré-processamento dos dados e geração dos modelos de predição de evasão foi utilizada a ferramenta *Google Colaboratory* (Colab), baseado no *Jupter Notebook*, com a linguagem *Python*, e as bibliotecas *Pandas* e *SKlearn*. O Colab foi escolhido por ser um serviço em nuvem gratuito e colaborativo, facilitando o trabalho em equipe.

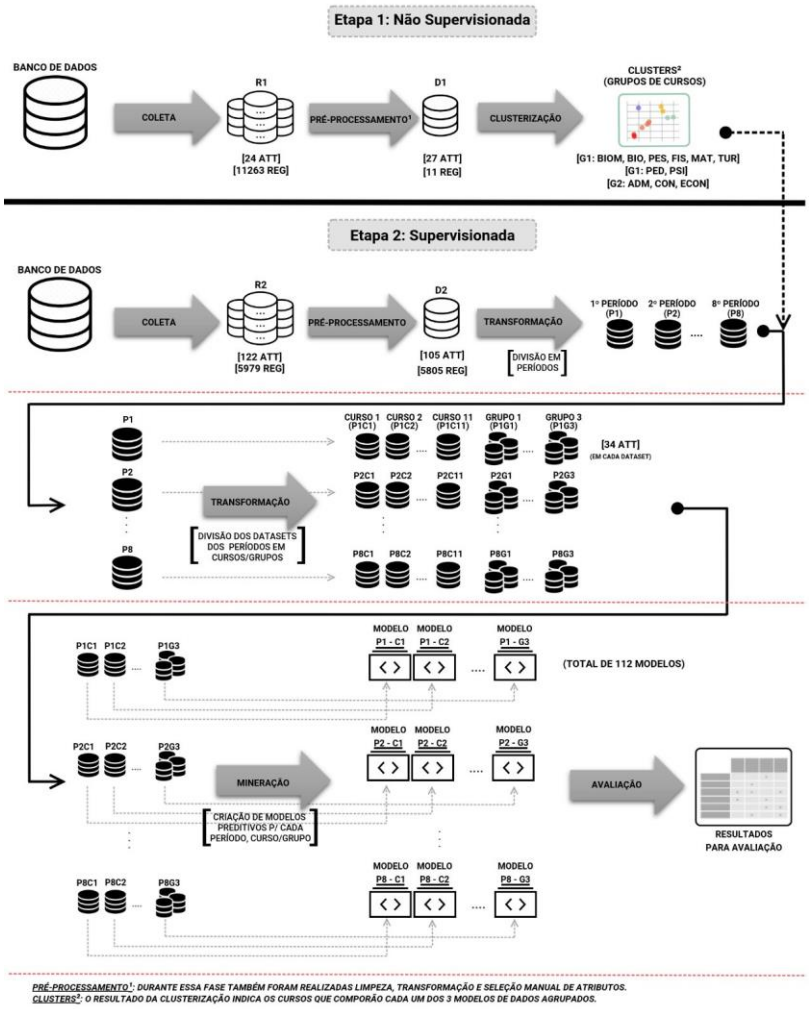

Figura 1 – Diagramação da metodologia proposta.

3.1 Etapa 1 – Não Supervisionada

Durante o pré-processamento, foram realizadas etapas de limpeza e seleção de atributos. Também foi conduzida a etapa de transformação, que envolveu a criação de atributos, conversão de atributos categóricos em numéricos e normalização dos dados. Como exemplo de limpeza, adotamos o método de imputação pela média para preencher valores ausentes. Em seguida, procedemos com uma seleção manual de atributos, excluindo identificadores e utilizando a análise do Coeficiente de Correlação de Pearson, na qual foram removidos os atributos com valores superiores a 0.95, resultando no Dataset de Trabalho 1 (D1) para a primeira etapa, com 27 atributos e 11 registros/tuplas (um por curso). Essa abordagem assegura que o resultado da clusterização seja composto por grupos de cursos semelhantes.

O conjunto de atributos é composto por 12 relacionados às taxas de evasão dos cursos (de 2012 a 2023), 9 baseados em médias, 1 relativo ao percentual de mulheres no curso, 2 de totais dos dados dos discentes para cada curso e 3 de características próprias

do curso (Área CNPQ, Grau Acadêmico e Turno). Para a construção desses atributos, excetuados os relativos às taxas de evasão, foram utilizados apenas dados relativos aos não evadidos (concluídos) e evadidos. Os atributos baseados em médias foram obtidos através do cálculo de informações dos discentes de um determinado curso. Como exemplo, a média da idade de ingresso de todos os alunos de ADM corresponde ao atributo 'mediaidadeingresso' para o curso ADM. Da mesma forma, foram calculados atributos de média do nº de membros da família, das notas do Enem, do índice de rendimento acadêmico, do nº de trancamentos em turmas bem como das reprovações. Quanto aos atributos de totais, tem-se de reprovações e trancamentos no curso.

Em seguida, aplicamos o algoritmo de clusterização *K-Means* sobre D1. Foram utilizados os métodos de *Elbow* e Silhueta para identificar do número ideal de clusters (K). Em ambos, 3 e 5 foram os quantitativos mais adequados, sendo o resultado com 3 clusters sutilmente maior, conforme os valores aproximados de Silhueta para 2, 3, 4, 5 e 6 clusters respectivamente listados a seguir: 0.4043, 0.5043, 0.4336, 0.4825 e 0.4358.

Devido à pequena discrepância entre os resultados para 3 e 5 *clusters*, foram avaliados os grupos gerados em ambos os cenários e optamos por descartar a opção com 5 *clusters*, pois um dos grupos resultantes possuía um único curso, inviabilizando a solução de aumento de dados para o mesmo. Assim, obtivemos os 3 seguintes grupos de cursos formados: Grupo 1 (G1) com BIO, BIOM, PES, FIS, MAT e TUR; Grupo 2 (G2) com PED e PSI; e Grupo 3 (G3) com ADM, CON e ECON. Na Figura 2, tem-se a representação gráfica desse resultado após aplicação da técnica de Análise de Componentes Principais para simplificar a representação dos dados.

Na Figura 3, são comparados os quantitativos de amostras referentes à classe minoritária (evadidos) presentes nos conjuntos de dados formados por cada grupo com os de cada curso individualmente. Nela, tem-se a representação para G1, G2 e G3 e seus cursos correlacionados, com os valores em destaque sobre as linhas do grupo e do curso com mais amostras. Assim é possível observar o ganho de dados dessa classe para cada período proporcionado pelos agrupamentos.

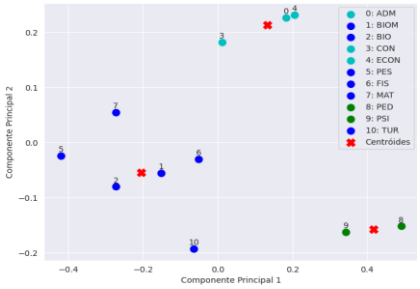

Figura 2 – Gráfico de dispersão com o resultado da clusterização (K=3).

Figura 3 – Total de evadidos por período e curso comparados aos dos agrupamentos.

Já na Tabela 1, apresenta-se a proporção do total de dados do grupo (e não apenas da classe minoritária) em relação ao do curso que o compõe, em valores percentuais. Para os cursos cujo benefício promovido pelo agrupamento foi mais discreto, observou-se um aumento de pelo menos 55%, como em PSI e 82% em CON. Todos os demais tiveram um ganho bastante superior, sendo a maioria acima de 400%.

3.2 Etapa 2 – Supervisionada

Após a coleta, foi iniciado o pré-processamento, com limpeza e seleção de atributos, seguido pela transformação dos dados. A Limpeza foi realizada conforme a Etapa 1, para os atributos coincidentes, embora a Etapa 2 tenha envolvido uma quantidade maior de características. Por exemplo, adotamos o método de imputação para preencher valores ausentes com valores mais frequentes para raça e estado civil.

Em seguida, realizamos a transformação gerando novos e ajustando os tipos de atributos. Foi realizada ainda uma seleção manual de atributos na qual excluiu-se atributos identificadores, utilizados para geração de outros, constantes e os com correlação maior do que 0.95. Como resultado, obteve-se em um *Dataset* de Trabalho 2 (D2) com 105 atributos e 5805 registros, sendo 3298 evadidos e 2507 não evadidos.

Tabela 1 – Proporção entre o total de dados do grupo e o curso que o compõe.

PER	BIO	BIOM	PES	FIS	MAT	TUR	PED	PSI	ADM	CON	ECON
1º	465,18%	587,62%	476,18%	566,98%	454,60%	475,24%	168,29%	59,42%	246,83%	100,77%	368,21%
2º	465,68%	510,98%	518,06%	486,81%	498,65%	523,83%	179,66%	55,66%	249,84%	93,30%	408,10%
3º	460,58%	496,42%	560,62%	450,00%	518,57%	528,57%	175,81%	56,88%	241,90%	90,02%	451,70%
4º	463,76%	477,51%	600,99%	431,42%	528,61%	526,76%	171,11%	58,44%	235,07%	89,85%	471,97%
5º	445,79%	464,83%	622,30%	418,13%	584,15%	516,83%	170,88%	58,52%	233,33%	85,59%	520,44%
6º	443,69%	437,08%	636,25%	403,42%	645,57%	520,00%	170,24%	58,74%	233,33%	85,35%	523,08%
7º	455,59%	409,01%	699,51%	386,35%	676,78%	509,29%	171,95%	58,16%	232,19%	84,29%	539,67%
8º	447,18%	398,08%	717,89%	384,11%	709,38%	509,41%	169,14%	59,12%	233,78%	82,73%	553,04%

Os atributos são apresentados na Tabela 2, incluindo informações socioeconômicas e acadêmicas dos estudantes e de seus cursos, divididos em duas categorias: atributos fixos, que não variam em todos os semestres; e informações acadêmicas cumulativas por semestre, que variam conforme o discente avança no curso. Para estes, foram gerados 8 atributos para cada tipo, a depender do período. Por exemplo, existem 8 atributos do tipo "Total de Aprovações", com terminações de '01' a '08', correspondente ao período referente, o qual indica o total de turmas aprovadas pelo discente até o final do semestre em questão. O atributo classe foi criado com base nos discentes que possuíam o status 'concluído' e os aptos a colarem Grau, classificados como 'Não Evadido'. Já como 'Evadido' foram classificados os discentes com status 'cancelado' e os que não efetuaram matrícula nos 3 últimos semestres (ou mais) até a coleta dos dados, seguindo as normas para cancelamento por abandono da instituição.

D2 foi então subdividido em 8 conjuntos de dados (P1 a P8), um para cada período do 1º ao 8º, sendo o último destinado a discentes no 8º período ou além. Cada conjunto inclui todos os atributos fixos e 10 cumulativos (correspondentes ao semestre em que o discente se encontra). Os dados também foram segmentados, de forma que todos os não evadidos (concluíram o curso) estejam presentes em todos os conjuntos de dados, exceto aqueles que concluíram antecipadamente. Por outro lado, os evadidos estão todos nos modelos do 1º período, mas o número cai com o avanço dos períodos à medida que os discentes evadem. Na Tabela 3, tem-se os totais de dados por período e classe.

Por fim, de cada conjunto de dados dos períodos, foram criados conjuntos para cada curso e agrupamento resultante da Etapa 1, totalizando 112, que geraram, cada um, modelos preditivos. Ou seja, 24 modelos referentes aos 8 períodos avaliados de cada um dos 3 grupos gerados, somados a 88 modelos (8 períodos dos 11 cursos) que são os modelos para comparação. O algoritmo aplicado foi o *Random Forest*, amplamente utilizado no contexto de evasão escolar (Dos Santos et al., 2021).

4. Resultados e Discussões

As métricas mais importantes para este estudo focam na qualidade da predição de que um discente não irá evadir, uma vez que erros nessa predição podem resultar na aplicação inadequada de políticas públicas preventivas. Foram utilizados, então, o Recall

(RC) e o Coeficiente de Correlação Matthews (Matthews Correlation Coefficient - MCC) para os modelos de predição de evasão, ambos obtidos por validação cruzada com 10 subdivisões. O RC é relevante em situações onde Falsos Negativos (discentes erroneamente previstos como Não Evadirão) são considerados mais prejudiciais do que Falsos Positivos (discentes previstos erroneamente como Evadirão). Quanto ao MCC, é uma métrica que avalia a qualidade das predições em modelos de classificação binária, considerando todas as quatro partes da matriz de confusão, sendo especialmente útil para lidar com conjuntos de dados desequilibrados, como os deste estudo.

Tabela 2 – Atributos dos modelos de predição.

Atributos Fixos	
Atributo	**Tipo**
Área CNPQ	Categórico
Evadido (target)	Boolean
Estado civil	Categórico
Grau do curso	Categórico
Idade de ingresso	Numérico
Intervalo em anos entre o ingresso e o ano de conclusão do ensino médio	Numérico
Notas nas provas do Enem (total de 5)	Numérico
Número da convocação em processo seletivo do SISU	Categórico
Opção de Inscrição no SISU	Categórico (1,2)
Período de ingresso	Categórico (1,2)
Raça	Categórico
Se concluiu em escola pública	Boolean
Se cursou componentes curriculares nos anos da pandemia	Boolean
Se possui crédito concedido ou crédito automático	Boolean
Se tem participação em atividades de ensino	Boolean
Sexo	Categórico (F, M)
Tipo de ação afirmativa de ingresso	Categórico
Total de membros da família	Numérico
Turno do curso	Categórico
UF de naturalidade	Categórico
Atributos Cumulativos	
Atributo	**Tipo**
Média acumulada	Numérico
Total de aprovações	Numérico
Total de aprovações em componentes obrigatórios no semestre	Numérico
Total de cancelamentos de turmas	Numérico
Total de reprovações em turmas obrigatórias no semestre	Numérico
Total de reprovações por falta	Numérico
Total de reprovações por falta e nota	Numérico
Total de reprovações por nota	Numérico
Total de trancamentos em turmas	Numérico
Total de trancamentos no curso	Numérico

Tabela 3 – Totais de dados presentes nos modelos de dados dos períodos.

	P1	P2	P3	P4	P5	P6	P7	P8
Evadidos	3298	2041	1560	1257	996	756	584	463
Não Evadidos	2507	2507	2507	2504	2504	2502	2498	2496

O valor do MCC varia de -1 a +1, sendo simétrico em relação às classes (trocar os positivos pelos negativos resultará no mesmo resultado). O coeficiente "+1" indica predição perfeita, "0" indica predição aleatória, e "-1" indica predição inversa (Chicco et al., 2021). Neste estudo, como não houve inversão das classes, considera-se a variação de 0 a 1, assim como para o RC. Chicco et al. (2020) afirmam que o MCC é uma métrica mais confiável para avaliar classificações binárias do que a Acurácia e o F1 Score. Eles destacam que, apesar da popularidade dessas métricas, em conjuntos de dados desequilibrados, podem ser enganosas, pois não consideram a proporção entre elementos positivos e negativos. Nas Tabelas 4 e 5 estão os resultados dessas métricas para cada período, curso e grupo referente.

4.1 Avaliação dos Resultados

Para avaliar estatisticamente a diferença entre os resultados de um modelo de um

curso e do grupo ao qual o curso integra em um mesmo período, foi aplicado o Teste de Hipótese Z com um nível de significância de 5%. A Tabela 6 mostra a avaliação sobre a métrica RC e MCC. Os valores "0", "+1" e "-1" representam os resultados em que não há diferença significativa, o modelo agrupado é significativamente melhor e o modelo agrupado é significativamente pior do que o do curso, respectivamente.

Tabela 4 – Resultados de RC e MCC para G1 e cursos que o compõem.

PER	BIO		BIOM		PES		FIS		MAT		TUR		G1	
	RC	**MCC**	**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**
1º	85,80%	59,84%	68,64%	56,76%	95,48%	70,85%	73,70%	64,49%	92,31%	54,54%	83,90%	46,08%	87,01%	66,22%
2º	82,00%	58,81%	66,48%	58,30%	92,25%	68,90%	71,71%	72,28%	85,89%	61,04%	74,80%	54,24%	79,93%	65,91%
3º	73,66%	66,00%	57,56%	55,37%	88,90%	65,39%	69,62%	69,19%	88,03%	68,35%	72,78%	60,43%	77,36%	69,21%
4º	68,63%	64,53%	56,79%	58,88%	88,16%	70,12%	73,45%	72,07%	84,88%	69,11%	77,86%	67,93%	78,15%	71,57%
5º	72,42%	64,98%	52,67%	51,64%	85,66%	75,00%	65,71%	71,58%	84,00%	70,60%	75,71%	68,05%	73,71%	69,42%
6º	65,18%	60,98%	49,17%	54,60%	86,92%	72,70%	58,33%	64,19%	79,64%	73,80%	73,22%	69,56%	74,84%	74,45%
7º	58,75%	63,52%	38,33%	47,17%	83,33%	79,45%	47,00%	60,34%	88,39%	83,93%	69,11%	65,43%	69,75%	71,11%
8º	63,81%	62,51%	53,33%	55,75%	85,00%	79,31%	48,33%	52,19%	81,0%%	77,24%	63,57%	60,67%	67,68%	71,78%

Tabela 5 – Resultados de RC e MCC para G2, G3 e cursos que os compõem.

PER	PED		PSI		G2		PER	ADM		CON		ECON		G3	
	RC	**MCC**	**RC**	**MCC**	**RC**	**MCC**		**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**	**RC**	**MCC**
1º	67,35%	65,69	66,41%	61,37%	61,76	60,23	**1º**	65,36%	51,58%	67,23%	57,16%	96,02%	72,56%	74,74%	58,48%
2º	53,00%	58,45	53,67%	49,09%	49,13	54,29	**2º**	49,82%	37,31%	52,89%	49,90%	87,03%	63,58%	64,87%	53,18%
3º	51,50%	62,57	53,64%	52,88%	49,62	54,31	**3º**	50,00%	42,79%	64,90%	64,16%	77,33%	54,38%	61,78%	57,41%
4º	60,00%	68,41	53,75%	51,64%	54,17	62,68	**4º**	39,05%	41,51%	61,36%	61,83%	75,71%	58,37%	59,48%	58,17%
5º	55,00%	67,02	52,86%	57,63%	54,89	63,38	**5º**	45,00%	48,45%	54,86%	63,13%	69,67%	63,19%	62,76%	62,07%
6º	55,00%	57,69	47,33%	56,18%	45,48	53,27	**6º**	41,50%	45,84%	60,00%	64,05%	73,00%	68,42%	60,48%	65,38%
7º	10%*	9,58%*	45,50%	54,99%	52,33	60,45	**7º**	48,33%	52,89%	64,33%	65,07%	81,50%	73,69%	64,17%	66,40%
8º	10%*	10%*	39,17%	49,43%	47,00	60,23	**8º**	41,67%	44,38%	56,00%	61,71%	77,50%	72,74%	56,73%	61,02%

Tabela 6 – Comparação dos resultados dos cursos e respectivos grupos com Teste Z.

(a) RC

PER	BIO	BIOM	PES	FIS	MAT	TUR	PED	PSI	ADM	CON	ECON
1º	0	+1	-1	+1	-1	+1	0	0	+1	+1	-1
2º	0	+1	-1	+1	-1	+1	0	0	+1	+1	-1
3º	0	+1	-1	+1	-1	0	0	0	+1	0	-1
4º	+1	+1	-1	+1	-1	0	0	0	+1	0	-1
5º	0	+1	-1	+1	-1	0	0	0	+1	+1	0
6º	+1	+1	-1	+1	0	0	-1	0	+1	0	-1
7º	+1	+1	-1	+1	-1	0	+1	+1	+1	0	-1
8º	0	+1	-1	+1	-1	0	+1	+1	+1	0	-1

(b) MCC

PER	BIO	BIOM	PES	FIS	MAT	TUR	PED	PSI	ADM	CON	ECON
1º	+1	+1	-1	0	+1	+1	0	0	+1	0	-1
2º	+1	+1	0	-1	+1	+1	0	0	+1	0	-1
3º	0	+1	0	0	0	+1	-1	0	+1	-1	0
4º	+1	+1	0	0	0	0	0	+1	+1	0	0
5º	0	+1	0	0	0	0	0	0	+1	0	0
6º	+1	+1	0	+1	0	0	0	0	+1	0	0
7º	+1	+1	-1	+1	-1	0	+1	0	+1	0	0
8º	+1	+1	-1	+1	0	+1	+1	+1	+1	0	-1

Analisando o resultado da Tabela 6a (RC), pode-se afirmar que o uso da solução proposta por este trabalho é viável para aproximadamente 72,73% (8/11) dos casos em todos os períodos, com exceção do 5º em que o percentual cresce para 81,82% (9/11). Já avaliando a viabilidade de uso da solução proposta pela Tabela 6b (MCC), podemos perceber resultados ainda melhores variando entre 81,82% (9/11) a 100% (11/11) dos modelos em 3 dos 8 períodos. Se avaliado pelo aspecto dos cursos, há 100% de viabilidade para BIO, BIOM, TUR, PSI e ADM em ambas as métricas. De forma geral, a proposta pode ser aplicada em aproximadamente 73,86% dos casos, obtendo uma melhora significativa em 40,91% aproximadamente, quando avaliados pelo RC, e em 88,64%, obtendo melhora significativa em 39,77% aproximadamente, quando avaliados

pelo MCC. Os casos em que as métricas de desempenho não exibem diferenças significativas, entendemos como positiva a aplicação da técnica, pois dados em maior quantidade podem ser ativos mais valiosos para análises preditivas (Fortuny et al., 2013). Estes, por sua vez, proporcionam uma representação mais abrangente da complexidade do problema, capturando padrões que podem não ser percebidos em amostras menores, sendo mais robustos e com uma melhor capacidade de generalização. Além disso, modelos com mais dados tendem a manter seu desempenho.

No caso de PED, há um relevante ganho da solução proposta demonstrado no 7º e 8º períodos. O aprendizado de máquina depende da disponibilidade de dados, e a ausência de dados suficientes da classe minoritária nos modelos individuais resulta em um desempenho inadequado na realização de predições precisas. Especificamente, esse cenário é representado com '*' na Tabela 5, mas também se reflete nos baixos índices do 8º de PSI, 4º de ADM e 7º de BIOM para RC. No entanto, ao adotarmos a estratégia de agrupamento e o subsequente aumento nos dados disponíveis, a capacidade de aprendizado e predição se expandem para casos em que anteriormente não eram viáveis. Além disso, há uma tendência de melhoria nos índices de casos em que as previsões não eram confiáveis devido aos baixos índices iniciais.

Em resumo, nosso método proposto mostrou eficácia significativa em prever a evasão em uma ampla gama de cursos, com melhorias particularmente notáveis em áreas com dados limitados, como PED, PSI, ADM e BIOM. Essencialmente, a importância do tamanho das amostras na aprendizagem de máquina fica clara ao observarmos esse progresso. O aumento na quantidade de dados desempenha um papel fundamental na capacidade do modelo de prever com precisão.

5. Conclusão

Neste trabalho, conduziu-se um processo de clusterização para identificar cursos com maior similaridade em uma IFES recém-criada, que dispõe de uma quantidade limitada de dados para estudos de MDE. O objetivo foi agrupar os dados desses cursos, gerando modelos mais robustos com uma quantidade maior de dados, de forma não sintética. Além disso, realizou-se um processo de classificação para predição de evasão, aplicando uma separação dos dados por períodos e cursos. Essa abordagem resultou na criação de modelos de predição a partir do agrupamento de dados dos cursos indicados pela clusterização, e na construção de modelos de predição para os cursos individualmente, possibilitando comparações entre eles através do teste de Hipótese Z.

Os resultados indicam a viabilidade de se aplicar a solução proposta em aproximadamente 73.86% dos casos (65/88) considerando o RC e 88.64% (78/88), MCC. Além disso, para o curso de PED, dentre os modelos do 7º e 8º períodos, o modelo agrupado é a única solução possível, dada a insuficiência de dados da classe minoritária para alcançar resultados superiores aos aleatórios. Esses resultados são significativos, proporcionando uma melhoria na identificação da evasão, o que possibilita uma gestão mais eficaz na aplicação de políticas públicas preventivas.

Resultados promissores foram alcançados neste estudo. No entanto, as métricas aferidas podem ser aprimoradas, considerando que o foco era a comparação entre os modelos agrupados e os individuais. Um refinamento em alguns processos do KDD, como a seleção automática e a transformação de atributos, pode eventualmente agregar valor aos modelos, como a criação de atributos com informações de discentes evadidos e não evadidos de forma dissociada (um atributo para cada classe). Outros algoritmos de classificação e clusterização podem ser testados, assim como seus parâmetros. Além disso, a etapa não supervisionada abordou os dados como um conjunto único, sem segmentação por períodos. Pode-se, futuramente, realizar a mesma segmentação, gerando 8 resultados de grupos e aplicando-os na classificação.

6. Referências

Baker, R; Isotani, S; Carvalho, A. **Mineração de dados educacionais: Oportunidades para o Brasil.** Revista Brasileira de Informática na Educação, v. 19, n. 02, p. 03, 2011.

Chicco, D; Jurman, G. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. BMC genomics, v. 21, n. 1, p. 1-13, 2020.

Chicco, D; Warrens, M. J; Jurman, G. **The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment**. IEEE Access, v. 9, p. 78368-78381, 2021.

De Brito, B. C. P; De Mello, R. F. L; Alves, G. **Identificação de atributos relevantes na evasão no ensino superior público brasileiro**. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação. SBC, 2020. p. 1032-1041.

Dos Santos, V. H. B; Saraiva, D. V; De Oliveira, C. T. **Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar**. In: Anais do XXXII Simpósio Brasileiro de Informática na Educação. SBC, 2021. p. 1196-1210.

Dutra, J. F; Souza, J. P. L. de; Fernandes, D. Y. de S. **Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do IFPB.** Revista Principia - Divulgação Científica e Tecnológica do IFPB, João Pessoa, v. 59, n. 3, p. 1009-1027, 2022.

Fayyad, U; Piatetsky-Shapiro, G; Smyth, P. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37-37, 1996.

INEP (2022). **Indicadores de fluxo da educação superior. 2021.** Disponível em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior. Acesso em 15/06/2023.

Fortuny, E. J; Martens, D; Provost, F. **Predictive modeling with big data: Is bigger really better?** Big Data, v. 1, n. 4, p. 215–226, 2013. PMID: 27447254.

Manhães, L. M. B; Cruz, S. **Predição do desempenho acadêmico de alunos da graduação utilizando mineração de dados**. XIX Simpósio de Pesquisa Operacional e Logística da Marinha. Rio de Janeiro, RJ, Brasil, v. 6, 2020.

Manhães, L. M. B; Cruz, S; Zimbrão, G. **Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais.** RJ: UFRJ, 2015.

Marques, L. T; Queiroz, P. G. G; De Castro, A. F; Marques, B. T; Silva, J. C. P. **Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura.** RENOTE 2019, 17, 194-203, 2019.

Romero, C; Ventura, S. **Educational data mining: a review of the state of the art.** IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40(6):601–618, 2010.

Santos, C. H. DC; Martins, S. de L; Plastino, A. **É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?** In: Anais do XXXII Simpósio Brasileiro de Informática na Educação. SBC, 2021. p. 792-802.

Saraiva, D. V; Pereira, S. S; Braga, R. B; de Oliveira, C. T. **Análise de Agrupamentos para Caracterização de Indicadores de Evasão**. In Anais do XXIX Workshop sobre Educação em Computação (pp. 238-247). SBC, 2021.

Viana, F. S; Santana, A. M; Rabêlo, R. de A. L. **Avaliação de Classificadores para Predição de Evasão no Ensino Superior Utilizando Janela Semestral**. In: Anais do XXXIII Simpósio Brasileiro de Informática na Educação. SBC, 2022. P. 908-919.