Avaliação de Classificadores para Predição de Evasão no Ensino Superior Utilizando Janela Semestral

Filipe Soares Viana[1], **André Macêdo Santana**[1], **Ricardo de Andrade Lira Rabêlo**[1]

[1]Departamento de Computação - Universidade Federal do Piauí (UFPI)
Bairro Ininga - CEP: 64049-550 - Teresina - PI - Brazil

{filipe,andremacedo,ricardoalr}@ufpi.edu.br

Abstract. *The high dropout rate in Federal Institutions of Higher Education (IFES) is a phenomenon that increasingly concerns teachers and educational managers. In this sense, strategies such as Educational Data Mining have been used to mitigate this problem. A proposal for classifying dropouts is presented in this work, generating trained models for each school semester. The experiments were carried out with data from an IFES in the Northeast of Brazil, on Computer Science and Information Systems courses. The results of the Mining process are promising with accuracy between 85% and 96% .*

Resumo. *O alto índice de evasão nas Instituições Federais de Ensino Superior (IFES) é um fenômeno que, cada vez mais, vêm preocupando docentes e gestores educacionais. Neste sentido, estratégias como Mineração de Dados Educacionais vêm sendo usadas para mitigar esse problema. Uma proposta para classificação de alunos entre Evadidos e Graduados é apresentada nesse trabalho, gerando modelos treinados para cada janela semestral. Os experimentos foram realizados com dados dos cursos de Computação e Sistemas de Informação da Universidade Federal do Piauí (UFPI). Os resultados do processo de Mineração são promissores com acurácias entre 85% e 96%.*

1. Introdução

A evasão estudantil é definida pelo Ministério da Educação (MEC) como a saída definitiva do estudante do curso de origem, sem concluí-lo. É um problema recorrente no Brasil e em outros países, tendo como consequências a formação de profissionais abaixo da capacidade, além de causar frustação pessoal aos envolvidos e um significante desperdício de recurso público. Considerando apenas a rede pública de ensino superior, a taxa de evasão em cursos de graduação é próxima de 40% [BRASIL 2020]. Já para os cursos de Ciência da Computação e Sistemas de Informação essa taxa está em torno de 70% [Hoed 2016]. Ao se analisar dados recentes do INEP em relação à evasão, 43% dos ingressantes de 2011 em curso de graduação na esfera federal se formaram, com taxa de 55% de desistência acumulada [BRASIL 2020].

Estimar o risco de evasão de aluno é uma tarefa fundamental, visto que prevê uma evasão futura e, assim, possibilita o desenvolvimento de melhores estratégias de permanência e êxito. Nesse sentido, desenvolver técnicas e metodologias que possam auxiliar a gestão na identificação automática de alunos com tendência a abandono dos cursos, por exemplo, é um desafio presente. Com o avanço constante das pesquisas em análise de

DOI: 10.5753/sbie.2022.224764

dados nos últimos anos, várias áreas de estudos específicas foram sendo desenvolvidas, dentre elas a Mineração de Dados Educacionais (MDE).

A MDE é uma combinação de conceitos entre banco de dados, estatística e aprendizado de máquina como método para explorar todo e qualquer conjunto de dados educacionais, de modo a absorver análises favoráveis para o norteamento de uma instituição no gerenciamento de suas ações e recursos [Fonseca et al. 2019]. Podemos então explicar ocorrências educacionais e assim viabilizar melhorias no processo do ensino e aprendizagem a partir de técnicas de mineração de dados na educação, melhorando o entendimento de como os estudantes aprendem [Romero and Ventura 2013].

Esse trabalho descreve um processo completo de mineração dos dados da Universidade Federal do Piauí para classificar os discentes e predizer entre evadido e graduado. Essa predição ocorre com modelos especializados para cada período do discente no curso, considerando assim a temporalidade dos dados. Com isso, os padrões e especificidades de cada momento do curso são considerados separadamente, gerando classificações granularizadas em períodos.

2. Trabalhos Relacionados

Todo o processo de mineração de dados é comumente focado na aplicação de métodos de aprendizagem de máquina. A mineração de dados educacionais em específico pode se valer de vários tipos como previsão, classificação, associações e agrupamento [Araujo 2014] .Por exemplo, pode-se auxiliar uma instituição a prever o desempenho de alunos com base no comportamento até aquele momento do curso, ou utilizar classificação para separar alunos em grupos com características em comum, de acordo com as necessidades. Para cada tipo de mineração de dados, encontramos vários algoritmos disponíveis, mas sem definição simples de qual utilizar para determinado tipo de análise. Com isso, cada trabalho geralmente utiliza de diversos métodos e então determinam o que melhor se encaixa no problema.

Em um estudo de revisão sistemática da literatura realizada por [Agrusti et al. 2019], verificou-se que dos 73 trabalhos de mineração de dados na previsão de evasão estudantil, 67% usaram o classificador Árvore de Decisão, seguido pela classificação bayesiana com 49%. Outro mapeamento sistemático foi realizado por [Torres Marques et al. 2019], verificando que técnicas de classificação vêm sendo altamente utilizadas na detecção de evasão escolar devido à alta precisão nas previsões, listando como os algoritmos mais utilizados: naive bayes (NB), support-vector machine (SVM), network of radial basis function (RBFNetwork), multilayer perceptron (MLP), k-nearest neigh-bours (IBk), Jrip, OneR, J48, PART e AdaBoostM1. Além disso, verificou-se também as ferramentas de mineração de dados mais utilizadas como sendo o Weka [1] no quesito evasão.

[Manhães 2020] objetivou em sua pesquisa utilizar técnicas de mineração de dados educacionais e detectar o risco de evasão estudantil, contribuindo com os gestores acadêmicos na tentativa de reduzir os índices de evasão nos cursos de graduação, decorrente de ações direcionadas aos alunos classificados como evadidos. Todo o processo analisou diversos cenários na instituição, dividindo o estudo em 7 partes. Os 6 primeiro

[1] https://www.cs.waikato.ac.nz/ml/weka/

estudos classificaram, cada um, discentes de um curso de graduação. O último estudo finaliza com análise de todos os cursos de graduação da instituição. Dependendo do curso analisado, a acurácia chegou a 96%.

Um trabalho onde o foco é analisar fatores de evasão, sem o intuito somente de detectar evasão, foi o proposto por [SANTOS 2020]. O trabalho mostra um resultado final de acurácia de 98,6% na detecção de um perfil evasivo, com coeficiente Kappa de 0,97. Além disso, o trabalho de Santos responde diversas perguntas em conjunto com uma análise direta dos dados, como por exemplo, se a quantidade de matérias selecionadas influencia o desempenho do aluno. Destaca-se, então, a importância também do conhecimento gerado pela classificação, e não somente o resultado inicial de acurácia entregue pelos algoritmos.

Um processo para predição de evasão em cursos técnicos no Instituto Federal de Santa Catarina foi executado por [Bitencourt and Ferrero 2019], atingindo acurácia de 86.3%. O modelo treinado é utilizado e consegue recuperar 25% dos potenciais discentes que iriam evadir, validando assim a importância da ferramente para a gestão.

O trabalho de [Santos et al. 2021] é um dos poucos que utiliza a semestralidade no estudo realizado. São realizados treinamentos e validações para os 10 primeiros semestres do curso, atingindo acurácia entre 79,31% e 98,25% utilizando o algoritmo de árvore de decisão. O sistema se propõe, a cada semestre, verificar se evadirá ou não baseado no desempenho do aluno nas disciplinas cursadas até aquele semestre.

Modelos de predição de evasão de discentes são construídos por [Filho et al. 2020] em seu trabalho, utilizando dados dos campi da instituição. Os autores elencam os melhores resultados obtidos para cada campus, onde percebe-se que existe uma variação de desempenho dependendo do campus analisado. Por exemplo, enquanto um campus gerou uma área da curva ROC de 0,55, outro atingia 0,81. Essa variação demonstra que os atributos e a estratégia utilizada não se mostrou genérica, possuindo melhor atuação em determinados conjuntos de dados. Também é realizado um estudo dos atributos mais importantes para cada campus, verificando nesse caso também a diferença para cada atributo entre os campi. Assim, enquanto em um campus o atributo "matricula retida" possui a maior importância, esse mesmo atributo em outro campus já possui importância ínfima.

Diferentemente da maioria dos trabalhos estudados, esse se propõe a montar modelos preditivos considerando o semestre do curso que o discente se encontra. Considerando que a realidade e situação do curso e do aluno pode mudar dependendo do momento que se encontra no curso, se mostra importante separar modelos preditivos para cada semestre. Tal abordagem torna a execução mais factível em ambiente real, podendo ser facilmente implementando na instituição.

3. Materiais e Métodos

Este trabalho utiliza como base o processo de descoberta de conhecimento em banco de dados, conhecido como *Knowledge Discovery in Databases* (KDD), abordado no trabalho de [Fayyad et al. 1996]. O autor define o KDD como um processo que identifica padrões validos, úteis e compreensíveis nos dados a serem analisados, realizando um processo que abrange todos os passos necessário desde a coleta até a obtenção dos resultados. O

KDD possui 5 fases: coleta, pré-processamento, transformação, mineração e avaliação dos resultados com as devidas interpretações. Alguns trabalhos, como o de Garcia (2016) e Fen (2021), tratam a transformação como parte do pré-processamento, não se atendo rigidamente às rotulações de etapas. Vale ressaltar que tal processo não possui um rigor na sua execução sequencial, é possível rever cada etapa a qualquer momento no processo à procura do melhor resultado.

Alunos são classificados entre discentes Evadidos e Graduados. Evadidos são todos aqueles que se desconectaram do curso, sem considerar os casos de trancamento do curso. Todos os que concluíram o curso são considerados como Graduados. Todos os dados nesse trabalho foram coletados com assistência de analistas de negócio da instituição e contato direto com analistas do setor de Tecnologia da Informação. Os dados são relativos aos discentes com ano de ingresso a partir de 2012 até 2020 do curso de Ciências da Computação. No total foram coletadas 727 instâncias, sendo 287 classificados como Evadidos, 92 como Graduados e 348 ainda ativos. O ano mínimo foi limitado em 2012 por se referir ao momento de início do Sistema de Seleção Unificada (SISU), ao mesmo tempo que ocorreu a implantação do sistema de gestão acadêmica na Instituição, evitando assim inconsistências na migração de dados anteriores.

A coleta ocorreu em consulta direta em banco de dados de um total de 134 atributos. Foram removidos atributos com valores únicos como nomes, identificadores (RG, CPF, etc) e similares (chaves do banco, log), além de atributos com mais que 50% de instâncias com valores desconhecidos. Em seguida, com ajuda de analistas da instituição, foram selecionados inicialmente 12 atributos referentes ao momento ingresso na instituição. Os atributos selecionados são detalhados na Tabela 1.

Tabela 1. Atributos sociais resultantes da seleção manual.

	Atributo	**Descrição**	**Tipo**
1	Ano_ingresso	Ano de ingresso do discente.	Número
2	Período_ingresso	Período de ingresso do discente (1 ou 2)	Categórico
3	Conclusão_escola_publica	Se o discente concluiu ensino médio na escola pública	Boolean
4	uf_naturalidade	É a UF da naturalidade do discente.	Categórico
5	sexo	Sexo do discente (M, F) transformados em números 1 e 0	Categórico
6	raca	Identificador de raça do discente.	Categórico
7	estado_civil	Identificador de estado civil	Categórico
8	Is_capital	Se o município de contato é a capital, sede da instituição	Categórico
9	Uf_contato	UF do contato do discente	Categórico
10	Cota	Informa se foi selecionado por cota na instituição	Boolean
11	tipo_acao_afirmativa	Identificador do tipo de ação afirmativa do qual o discente participa, se existente.	Categórico
12	idade	idade no momento de ingresso no curso	Numérico

Outros atributos selecionados foram os de caráter acadêmico, derivados de dados dos discentes diretamente do banco de dados, como notas por período e os registros de cada desciplina matriculada. Os dados do curso analisado são separados em 6 conjuntos referentes aos 6 primeiros períodos do discente, cada um com dados classificados em Evadidos e Graduados. O total de atributos para cada período do discente totaliza 26 atributos, sendo 12 atributos de ingresso e 11 acadêmicos. O limite é definido no 6º período pois a partir do 7º período ocorre uma perda relevante no quantitativo de discentes evadidos, acarretando em perda no desempenho dos algoritmos de aprendizagem. O conjunto de ativos dos discentes não são considerados para o treinamento, sendo posteriormente utilizados em uma avaliação de testes.

Os atributos acadêmicos possuem tanto atributos específicos para cada modelo por período quanto atributos gerais. Os atributos especializados são atributos que possuem vertentes para cada período treinado. Por exemplo, o atributo de Quantidade de Reprovação por Falta (quant_repr_falta) possui variações para cada modelo, dos períodos 1 ao 6. Assim, são gerados os atributos quant_repr_falta_1, utilizado no treinamento do 1º período, quant_repr_falta_2 no 2º período e assim sucessivamente, até o 6º período. Cada um desses atributos possui carater cumulativo, ou seja, o atributo de determinado período representa o valor acumulado de todos os períodos anteriores até o período em questão. A Tabela 2 exibe os atributos especializados por período e os não especializados.

Tabela 2. Atributos acadêmicos.

	Atributo	**Descrição**	**Especialista**
1	quant_disciplina_concluiu_prazo (qdcp)	Quantidade de disciplinas concluídas no prazo.	Sim
2	quant_repr_falta (qrf)	Quantidade de reprovação por falta.	Sim
3	quant_repr_nota (qrn)	Quantidade de reprovação por nota.	Sim
4	media_notas (mn)	Média das nota quando APROVADO.	Sim
5	media_notas_reprovado (mnr)	Média das notas quando o discente REPROVA.	Sim
6	quant_aprovadas_disciplinas_ef (qade)	Quantitativo de disciplinas APROVADO via Exame Final.	Sim
7	quant_trancamento_disciplinas (qtd)	Quantitativo de disciplinas trancadas.	Sim
8	quant_cancelamento_disciplinas (qcd)	Quantitativo de disciplinas canceladas.	Sim
9	se_possui_cancelamento_disciplina (spcd)	Sim/Não possui cancelamento de disciplina.	Sim
10	possui_graduacao_concluida (pgc)	Se possui graduação concluída anteriormente.	Não
11	possui_evasao_anterior (pea)	Se possui evasão em alguma graduação anteriormente.	Não

4. Resultados

Cada um dos conjuntos de dados temporais referentes ao período passa pelo processo de seleção de atributos. Para isso, é realizada inicialmente uma análise para definir o quantitativo de atributos que seriam selecionados em cada procedimento e qual algoritmo utilizar. Para a decisão, foi utilizado o algoritmo *Random Forest* (RF), por ser um algoritmo amplamente utilizado pela literatura, e executado o treinamento e teste com validação cruzada de 5 grupos. A acurácia foi calculada para cada algoritmo de seleção, sendo eles qui-quadrado (CHI2), *Mutual Information* (MI) [Rajab and Wang 2020] e análise de variância (ANOVA) [Yakub et al. 2016]. Cada execução foi realizada utilizado a seleção de 5, 10, 15, 20 e 25. Com isso, constatou-se que o melhor ganho foi na utilização do CHI2 selecionado os 15 primeiros atributos, como visualizado na Tabela 3. O CHI2 possui a maior faixa de acurácia, principalmente para seleção de 15 atributos, em especial no 1º período.

Tabela 3. Acurácia por período na execução de algoritmos de seleção de atributos para cada quantitativo de atributos selecionados.

	CHI2					**MI**					**ANOVA**				
Per	**5**	**10**	**15**	**20**	**25**	**5**	**10**	**15**	**20**	**25**	**5**	**10**	**15**	**20**	**25**
1	78%	81%	**85%**	82%	82%	80%	83%	83%	83%	83%	78%	82%	83%	81%	81%
2	84%	**87%**	**87%**	86%	**87%**	85%	**87%**	86%	85%	**87%**	85%	84%	85%	**87%**	**87%**
3	86%	88%	**89%**	88%	88%	85%	**90%**	**90%**	87%	**90%**	87%	87%	88%	88%	88%
4	86%	90%	**92%**	91%	91%	88%	88%	91%	**92%**	91%	91%	90%	91%	90%	90%
5	91%	**94%**	93%	93%	92%	90%	92%	92%	91%	92%	92%	91%	93%	92%	92%
6	94%	94%	**95%**	94%	**95%**	93%	94%	94%	94%	94%	93%	93%	94%	94%	94%

Ao executar a seleção de atributos para cada período, pode-se verificar os atributos mais apropriados para a classificação desejada (Tabela 4). A Figura **??** exibe os 15

atributos mais relevantes filtrados para o curso de Computação para o 3º período. De forma geral, os atributos acadêmico predominam como os atributos com maior nível de importância selecionados, situados sempre nas primeiras colocações do ranqueamento obtido.

Tabela 4. Ranking para cada período dos atributos selecionados com CHI2.

	P1	P2	P3P	P4	P5	P6
quant_disciplinas_concluiu_prazo	1°	1°	1°	1°	1°	2°
media_notas,	2°	7°	7°	10°	2°	1°
quant_repr_falta_semestre	3°	2°	2°	2°	3°	3°
quant_cancelamento_disciplinas	4°	6°	6°	4°	4°	4°
possui_evasao_anterior_na_ufpi	5°	3°	3°	3°	5°	5°
media_notas_reprovado	6°	-	-	11°	-	-
quant_repr_nota_semestre	7°	4°	4°	5°	6°	6°
quant_aprovadas_disciplinas_ef	8°	5°	5°	6°	8°	9°
id_uf_naturalidade_99	9°	9°	10°	8°	-	-
idade	10°	8°	9°	9°	10°	14°
quant_trancamento_disciplinas	11°	-	-	-	-	15°
id_tipo_acao_afirmativa_4.0	12°	10°	8°	7°	7°	7°
raca_Negro	13°	11°	11°	-	9°	8°
uf_contato_tipo_21	14°	-	-	14°	13°	-
cota	15°	-	-	13°	12°	11°
id_tipo_acao_afirmativa_6.0	-	12°	15°	15°	14°	13°
id_estado_civil_2.0	-	13°	12°	12°	11°	10°
periodo_ingresso_1	-	14°	13°	-	-	-
periodo_ingresso_2	-	15°	14°	-	-	-

Os atributos de Média de Notas, quantidade de disciplinas concluídas no prazo, quantidade de disciplinas canceladas, quantidade de reprovação por falta, quantidade de reprovação por nota e quantidade de disciplinas aprovadas por exame final estão presentes constantemente entre as 10 primeiras colocações do raqueamento, com algumas variações de importância entre os períodos (Figura 1). Por exemplo, o atributo média de notas possui uma maior importância no primeiro e no sexto período, enquanto perde nos outros conjuntos de dados, mesmo que ainda com grande relevância.

Para cada um desses conjuntos de dados gerados são executados algoritmos de aprendizagem de máquina, classificando os dados e gerando assim modelos especialistas para cada período. Assim, quando um discente, por exemplo, finaliza o 3º período, utiliza-se o modelo gerado pelo conjunto de dados treinado até o 3º período, possuindo assim uma classificação mais factível. É realizada uma comparação dos algoritmos de aprendizagem para classificar e, consequentemente, predizer a Evasão do discente. Os algoritmos utilizados são *Random Forest* (RF), *Decision Tree* (DT), *Extra Trees* (ET), *Multilayer Perception* (MLP), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e *Gaussian Naive Bayes* (GNB), todas utilizando a biblioteca Scikit-Learn [Pedregosa et al. 2011] para a linguagem Python. Por fim, cada classificação será validada utilizando Acurácia, Precisão, *Recall*, F1 score, Índice Kappa, área da curva ROC e os valores que constituem a matriz de confusão, sendo eles o Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) [GOMES 2002]. O resumo de todo o processo pode ser visualizado na Figura 2.

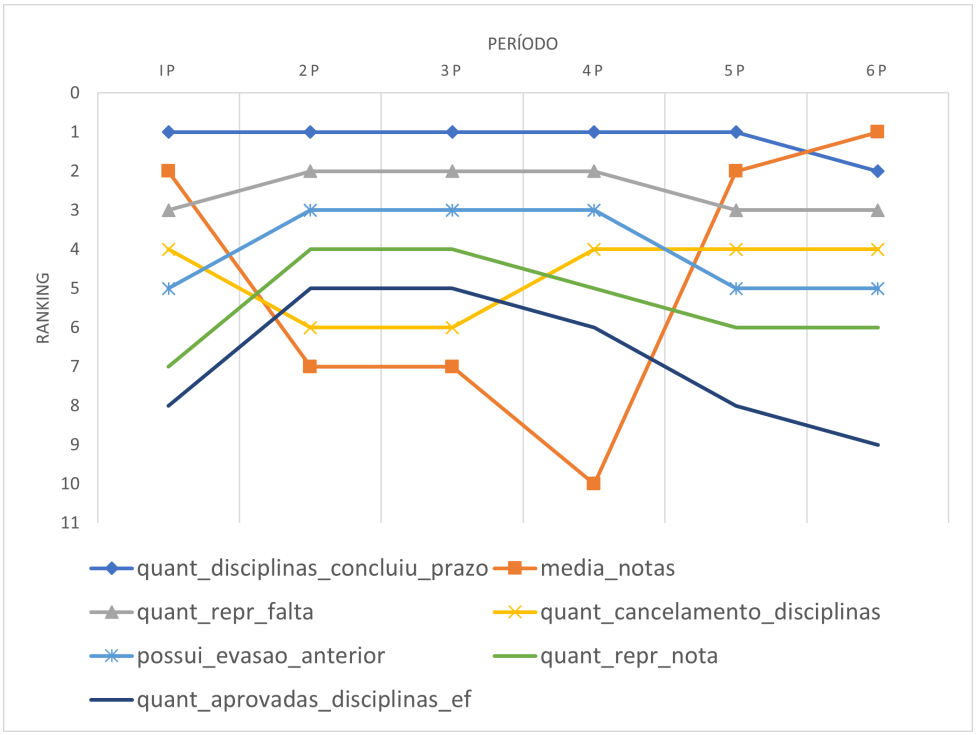

Figura 1. Os 7 primeiros atributos sempre presente nas 10 primeiras colocações

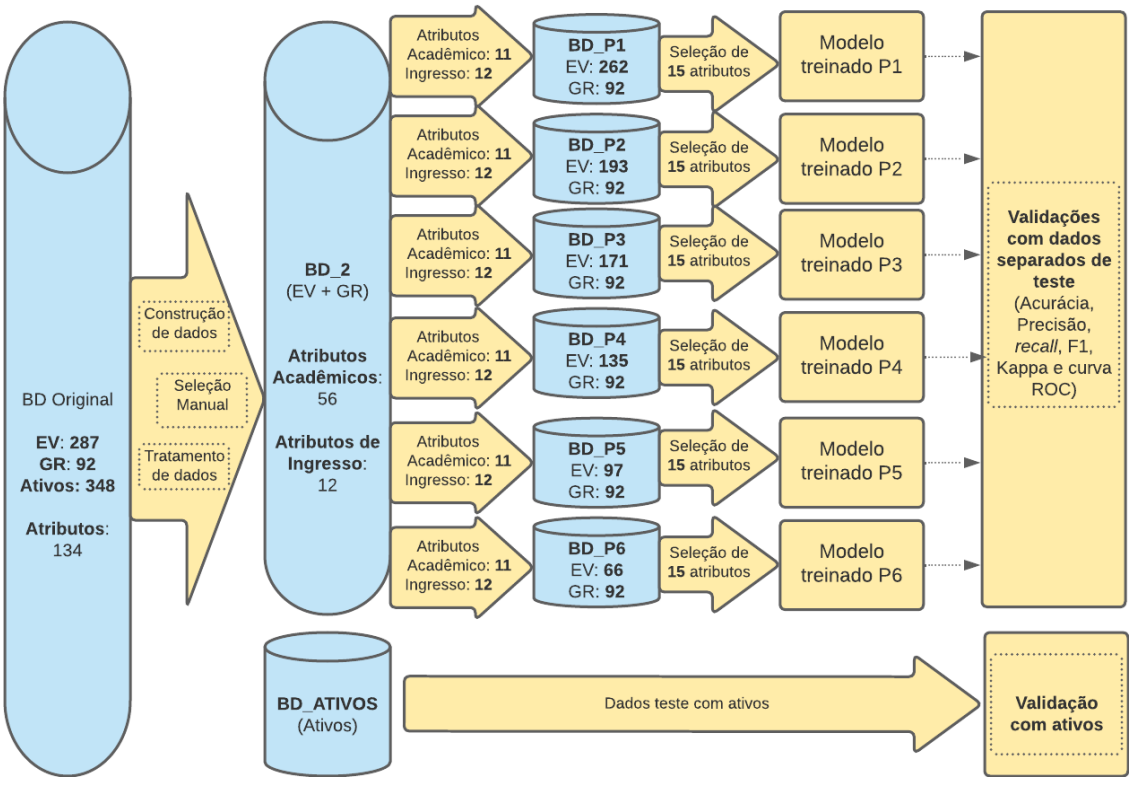

Figura 2. Processo geral

5. Resultados e Discussões

O primeiro experimento classifica os dados com ingressantes de 2012 a 2020. Foi executado o algoritmo CHI2 para cada período analisado, do 1º ao 6º, seguindo em cada um a execução dos algoritmos de aprendizagem. A Tabela 5 exibe os valores médios de acurácia obtidos para cada período por algoritmo de classificação.

Tabela 5. Acurácia obtidas por período para cada algoritmo utilizado.

Per	RF	DT	ET	MLP	SVM	NKK	GNB
1	**86,1%**	82,9%	83,5%	**86,7%**	85,2%	83,2%	81,4%
2	90,4%	85,4%	86,8%	88,2%	87,1%	83,6%	82,9%
3	**91,2%**	86,9%	**91,2%**	89,2%	**90,0%**	85,8%	87,3%
4	**91,6%**	88,5%	**92,0%**	**92,0%**	**91,1%**	**92,1%**	85,4%
5	**95,7%**	92,6%	**93,7%**	91,5%	91,5%	92,6%	88,9%
6	**94,3%**	91,2%	**95,6%**	**95,6%**	**94,3%**	93,7%	87,3%

Pode-se verificar que, de forma geral, todos possuem um bom desempenho, mas com um melhor desempenho no geral para os algoritmos RF, ET, MLP e SVM. Ao analisar as acurácias obtidas e realizando o teste estatístico de Friedman [Friedman 1937], constatou-se que não existe uma diferença significativa entre eles. A Tabela 6 exibe os valores de validação detalhados dos principais algoritmos, exibindo a média e desvio padrão de cada validação.

Os valores se mostram estáveis devido ao baixo desvio padrão nas médias obtidas. Os valores de Precisão de *Recall* estão sempre próximos dos valores de acurácia, atingindo índices Kappa superiores a 0,6, chegando a 0,8 a partir do 3º período, considerado um valor excelente de acordo com [Landis and Koch 1977]. Pode-se observar que, mesmo com a similaridade dos resultados entre RF e MLP, verifica-se que, dependendo do período, um algoritmo possui melhor desempenho que o outro. Por exemplo, o algoritmo RF atinge melhor resultado no 1º período, mas o ET e MLP possui melhores resultados no 4º período. Com isso, é possível separar o treinamento por período e elencar algoritmos para cada conjunto de dados analisado.

Uma segunda avaliação é proposta nesse trabalho com o objetivo de avaliar parcialmente o treinamento na tentativa de predizer a evasão em novos dados. Para isso, foi construída uma modelagem de treinamento com algoritmo *Random Forest*, utilizando os dados de Graduados (GR) e Evadidos (EV), para predizer os dados das instâncias que possuem status ativo (ainda não preditos). Após executar os testes, é gerado o quantitativo de dados preditos como evadidos e graduados a cada ano, com ingressantes de 2013 a 2020.

Inicialmente, são gerados 6 modelos diferentes, treinados para cada período do discente, com dados acadêmicos correspondentes. Para cada ano de ingresso, entre 2013 e 2020, são selecionados os discentes ativos e executados em seus respectivos modelos, com exceção dos períodos maiores que o 6º período. Ou seja, os discentes que estão no 1º período serão executados no modelo referente ao 1º período, e assim por diante. Para todos com período superior ao 6º, é utilizado o modelo treinado com os dados do 6º período. Após aplicar o modelo com os dados dos ativos, cada instância recebe seu devido rótulo, predizendo assim se irá evadir ou concluir o curso.

Tabela 6. Valores de avaliação detalhados para os algoritmos RF e MLP.

	Per	Acurácia	Recall	Precisão	F1 Score	Kappa	Curva ROC	TN	FP	FN	TP
RF	1	86,1% (0,03)	0,89 (0,03)	0,92 (0,03)	0,90 (0,02)	0,66 (0,08)	0,84 (0,05)	14	4	5	44
	2	90,4% (0,04)	0,93 (0,04)	0,93 (0,04)	0,93 (0,03)	0,78 (0,09)	0,89 (0,05)	16	3	3	35
	3	91,2% (0,05)	0,92 (0,09)	0,95 (0,02)	0,93 (0,04)	0,81 (0,09)	0,91 (0,03)	17	2	3	31
	4	91,6% (0,05)	0,92 (0,06)	0,94 (0,05)	0,93 (0,04)	0,83 (0,10)	0,92 (0,05)	17	2	2	25
	5	95,7% (0,04)	0,93 (0,07)	0,99 (0,02)	0,96 (0,04)	0,92 (0,09)	0,96 (0,04)	18	0	1	18
	6	94,3% (0,02)	0,91 (0,03)	0,96 (0,06)	0,93 (0,03)	0,88 (0,05)	0,94 (0,02)	18	1	1	12
ET	1	83,5% (0,02)	0,89 (0,02)	0,89 (0,02)	0,89 (0,02)	0,58 (0,06)	0,79 (0,03)	13	6	6	44
	2	86,8% (0,05)	0,89 (0,05)	0,91 (0,04)	0,90 (0,04)	0,70 (0,12)	0,85 (0,06)	15	3	4	34
	3	91,2% (0,03)	0,94 (0,02)	0,93 (0,03)	0,93 (0,02)	0,81 (0,06)	0,90 (0,03)	16	2	2	31
	4	92,0% (0,04)	0,92 (0,04)	0,95 (0,04)	0,93 (0,03)	0,84 (0,08)	0,92 (0,04)	17	1	2	25
	5	93,7% (0,05)	0,92 (0,05)	0,96 (0,04)	0,94 (0,05)	0,87 (0,09)	0,94 (0,05)	18	1	2	18
	6	95,6% (0,03)	0,93 (0,07)	0,97 (0,04)	0,95 (0,03)	0,91 (0,05)	0,95 (0,03)	18	0	1	12
MLP	1	86,7% (0,04)	0,90 (0,03)	0,91 (0,03)	0,91 (0,03)	0,67 (0,10)	0,84 (0,05)	14	4	5	45
	2	88,2% (0,02)	0,89 (0,02)	0,94 (0,03)	0,91 (0,01)	0,74 (0,05)	0,88 (0,03)	16	2	4	33
	3	89,2% (0,05)	0,90 (0,04)	0,93 (0,04)	0,92 (0,04)	0,77 (0,10)	0,89 (0,05)	16	2	3	30
	4	92,0% (0,05)	0,92 (0,06)	0,95 (0,05)	0,93 (0,04)	0,84 (0,10)	0,92 (0,05)	17	1	2	25
	5	91,5% (0,04)	0,92 (0,03)	0,92 (0,05)	0,92 (0,04)	0,83 (0,08)	0,92 (0,04)	17	2	2	18
	6	95,6% (0,04)	0,96 (0,04)	0,95 (0,09)	0,95 (0,04)	0,91 (0,08)	0,96 (0,03)	18	1	1	13
SVM	1	85,2% (0,05)	0,92 (0,05)	0,88 (0,03)	0,90 (0,03)	0,61 (0,12)	0,80 (0,06)	12	6	4	45
	2	87,1% (0,07)	0,88 (0,09)	0,93 (0,05)	0,90 (0,06)	0,72 (0,13)	0,87 (0,06)	16	3	4	33
	3	90,0% (0,03)	0,91 (0,02)	0,94 (0,05)	0,92 (0,02)	0,78 (0,06)	0,90 (0,04)	16	2	3	31
	4	91,1% (0,04)	0,91 (0,06)	0,94 (0,06)	0,92 (0,04)	0,82 (0,09)	0,91 (0,05)	17	2	2	24
	5	91,5% (0,05)	0,89 (0,06)	0,94 (0,03)	0,91 (0,05)	0,83 (0,09)	0,92 (0,05)	17	1	2	17
	6	94,3% (0,05)	0,91 (0,09)	0,96 (0,06)	0,93 (0,06)	0,88 (0,10)	0,94 (0,05)	18	1	1	12

Os valores dos ativos preditos são somados e representados na Tabela 7, com as devidas análises. A linha Real possui o quantitativo de discentes já classificados como EV e GR. A linha de Predição de Ativos (PA) exibe o quantitativo dos ativos preditos no ano, tanto como evadidos como graduados. Para cada ano é realizado o cálculo da proporção de evadidos naquele ano, em conjunto com os dados existentes, e os previstos.

Tabela 7. Quantitativo de evadidos e graduados preditos para cada anos entre 2013 e 2020, com cálculo de porcentagem de evadidos para cada ano, do curso de Computação.

	2013		2014		2015		2016		2017		2018		2019		2020	
	EV	**GR**	**EV**	**GR**	**EV**	**GR**	**EV**	**GR**	**EV**	**GR**	**EV**	**GR**	**EV**	**GR**	**EV**	**GR**
Real	50	26	50	21	42	10	35	6	21	0	20	0	4	0	5	0
PA	2	4	8	1	7	11	14	20	25	28	27	30	41	23	38	26
Total	52	30	58	22	49	21	49	26	46	28	47	30	45	23	43	26
Ev	**63,41%**		**72,50%**		**70,00%**		**65,33%**		**62,16%**		**61,04%**		**66,18%**		**62,32%**	

Os resultados mostram que os valores de proporção de evadidos para cada ano possui uma variação praticamente constante, com pouca variação. Importante observar que a taxa referente ao ano de 2013 reflete uma taxa real de proporção de evadidos, visto que a grande maioria dos dados são de valores reais. Os valores subsequentes, ao se manterem próximos, mostram que não existe uma fuga desse padrão.

6. Conclusão

Esse trabalho realizou o processo de coleta, pré-processamento, transformação, classificação e validação em dados de discentes do curso de Computação de uma instituição de ensino federal. A arquitetura separa os dados por período, do 1º ao 6º

período, com a proposta de criar um modelo treinado para cada. Cada modelo é construído com dados de ingresso fixos e dados acadêmicos cumulativos e variáveis. Ou seja, os dados acadêmicos mudam para cada modelo, pois são dados cumulativos até aquele momento do curso. Importante ressaltar que o modelo referente ao 6º período é utilizado como modelo para prever todos os discentes que concluam do 6º período em diante.

Para validar o processo, foram propostos 2 experimentos para cada curso e seus respectivos modelos. Um dos experimentos utiliza validação cruzada com 5 grupos para verificar, de forma geral, a acurácia e demais índices de validação dos dados. O segundo experimento propõe uma predição dos discentes ativos até o momento da coleta dos dados, treinando com dados de 2012 a 2020 e prevendo todos os ativos. Assim, é realizado uma análise onde verifica-se que, ao utilizar o predito com dados dos ativos, a taxa de evasão do curso se mantém, contribuindo assim para validação dos modelos.

Uma análise inicial exibe claramente que os atributos que mais são relevantes na clasificação de evasão de um discente são os acadêmicos, principalmente os referentes à média de nota em cada período. O quantitativo de disciplinas concluídas no prazo se mostra muitas vezes como a mais importante, considerando assim que discentes que sigam o curso sem irregularidades na grade de disciplinas possuem um padrão de quem provavelmente concluirá o curso.

Os resultados são bem promissores, com valores de acurácia superiores a 86% na validação cruzada. Esse valor sobe a cada período classificado, com valor ótimo no 5º período na maioria dos algoritmos, chegando à 95,7%. Os algoritmos possuem comportamentos diferentes para cada período, podendo considerar algoritmos específicos para cada. Enquanto o RF possui melhor acurácia quando gerado o modelo para o 1º e 2º período, o melhor algoritmo para o 4º e 6º período é o ET e MLP. Nesses casos, pode-se gerar uma arquitetura onde para cada período seja atribuído um modelo específico que possua maior taxa de acurácia.

Trabalhos futuros podem ser realizados com análise de outros cursos e o agrupamento de cursos para verificar se existe uma variação na aprendizagem, se é possível gerar modelos genéricos ou se modelos específicos são mais vantajosos. Importante também gerar outros testes com diferentes anos, validando assim se a taxa de acurácia se mantem.

Outras abordagens técnicas podem ser consideradas, como a utilização de outros algoritmos de seleção de atributos à procura de um ponto ótimo na seleção. Também é possível realizar variações na execução de cada algoritmo, visto que cada um possui uma gama de parâmetros para ajustar sua execução, o que pode variar para cada caso.

Referências

Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, pages 161–182 Pages.

Araujo, F. H. D. d. (2014). Descoberta de conhecimento em base de dados para o aprendizado da regulação médica/odontológica em operadora de plano de saúde. Master's thesis, Pós-Graduação em Ciência da Computação, Universidade Federal do Piauí.

Bitencourt, P. B. d. and Ferrero, C. (2019). Predição de risco de evasão de alunos usando métodos de aprendizado de máquina em cursos técnicos. In *Anais dos Workshops*

do VIII Congresso Brasileiro de Informática na Educação (CBIE 2019), page 149. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC).

BRASIL (2020). Instituto nacional de estudos e pesquisas educacionais anísio teixeira. Site do INEP.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

Filho, F., Vinuto, T., and Leal, B. (2020). Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1132–1141, Porto Alegre, RS, Brasil. SBC.

Fonseca, S. O. d., Namen, A. A., Moura Neto, F. D., Silva, A. D. R., Ortigão, M. I. R., and Rohrer, U. A. B. V. (2019). Mineração de dados orientada pelo domínio educacional: uma prova de conceito. *Estudos em Avaliação Educacional*, 30(74):420.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

GOMES, A. K. (2002). Classificadores simbólicos utilizando medidas de avaliação e de interessabilidade. Master's thesis, Pós-Graduação em Ciência da Computação e Matemática Computacional, Universidade de São Paulo.

Hoed, R. M. (2016). Análise de evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação. Master's thesis, Pós-Graduação em Computação Aplicada, Universidade de Brasília.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Manhães, L. M. B. (2020). Predição do desempenho acadêmico de alunos da graduação utilizando mineração de dados. *Simpósio de Pesquisa Operacional e Logística da Marinha - Publicação Online*, 3(1):2050 – 2064.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rajab, M. and Wang, D. (2020). Practical challenges and recommendations of filter methods for feature selection. *Journal of Information & Knowledge Management*, 19(01):2040019.

Romero, C. and Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27.

Santos, C. H. D. C., Martins, S. d. L., and Plastino, A. (2021). É possível prever evasão com base apenas no desempenho acadêmico? In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE 2021)*, page 792–802, Brasil. Sociedade Brasileira de Computação - SBC.

SANTOS, K. J. d. O. S. (2020). Education data mining para apoio à gestão estratégica da identificação de perfis evasivos e atenuação da evasão escolar no ensino superior. Mas-

ter's thesis, Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Sergipe.

Torres Marques, L., Félix De Castro, A., Torres Marques, B., Carvalho Pereira Silva, J., and Gabriel Gadelha Queiroz, P. (2019). Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. *RENOTE*, 17(3):194–203.

Yakub, S., Arowolo, M., Abdulsalam, S., and M.D., S. (2016). A feature selection based on one way anova for microarray data classification. pages 30–35.